VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA CHEMICKÁ
FACULTY OF CHEMISTRY

ÚSTAV FYZIKÁLNÍ A SPOTŘEBNÍ CHEMIE
INSTITUTE OF PHYSICAL AND APPLIED CHEMISTRY

ANALÝZA ZMĚN GENOMU C. NECATOR PO EVOLUČNÍ ADAPTACI
ANALYSIS OF C. NECATOR GENOME CHANGES AFTER EVOLUTIONARY ADAPTATION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Štěpán Kroupa**
AUTHOR

VEDOUCÍ PRÁCE — **doc. Mgr. Václav Brázda, Ph.D.**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA CHEMICKÁ

Zadání bakalářské práce

Číslo práce:	FCH-BAK1519/2019	Akademický rok:	2019/20
Ústav:	Ústav fyzikální a spotřební chemie		
Student:	**Štěpán Kroupa**		
Studijní program:	Chemie a chemické technologie		
Studijní obor:	Chemie pro medicínské aplikace		
Vedoucí práce:	**doc. Mgr. Václav Brázda, Ph.D.**		

Název bakalářské práce:

Analýza změn genomu C. necator po evoluční adaptaci

Zadání bakalářské práce:

Literární rešerše se zaměřením na bakteriální genomy, analýza next generation sekvenačních dat pomocí bioinformatických nástrojů služby galaxy,, analýza polymorfismů, genů– Vyhodnocení sekvenačních dat.

Termín odevzdání bakalářské práce: 31.7.2020:

Bakalářská práce se odevzdává v děkanem stanoveném počtu exemplářů na sekretariát ústavu. Toto zadání je součástí bakalářské práce.

Štěpán Kroupa student(ka)	doc. Mgr. Václav Brázda, Ph.D. vedoucí práce	prof. Ing. Miloslav Pekař, CSc. vedoucí ústavu

prof. Ing. Martin Weiter, Ph.D.
děkan

V Brně dne 31.1.2020

ABSTRAKT

Tato práce se zabývá analýzou mutačních změn bakteriálních populací *Cupriavidus necator* H16 po evoluci za různých stresových podmínek. Tato analýza byla provedena zpracováním dat z metody sekvenování genomu „Next Generation Sequencing“, která byla provedena externí společností DNALink. Na základě bioinformatického postupu byl sestaven seznam mutačních změn pro každou adaptovanou populaci. Dále byly stanovené mutační změny asociovány s konkrétními oblastmi referenčního genomu *Cupriavidus necator* H16 z NCBI a analyzovány dle dostupných informací. Na závěr byl diskutován případný vliv stanovených mutací na produkci zásobních polymerů polyhydroxyalkanoátů.

Klíčová slova: *Cupriavidus necator* H16, evoluční adaptace, stresové podmínky, polyhydroxyalkanoáty, next generation sequencing, analýza mutací, bioinformatika

ABSTRACT

This bachelor’s thesis deals with analysis of mutations in bacterial populations of *Cupriavidus necator* H16 evolved in distinct stress conditions. This analysis was performed by processing data from the genome sequencing method „Next Generation Sequencing“, outsourced through the company DNALink. A list of mutations for each adapted population was constructed through bioinformatic methods. These mutations were then associated with specific areas of the reference *Cupriavidus necator* H16 genome from NCBI and analysed according to available information. Finally, the effect of these mutations on production of storage polymers polyhydroxyalkanoates was discussed.

Key words: *Cupriavidus necator* H16, evolutionary adaptation, stress conditions, polyhydroxyalkanoates, next generation sequencing, analysis of mutations, bioinformatics

KROUPA, Štěpán. *Analýza změn genomu C. necator po evoluční adaptaci.* Brno, 2020. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/123816. Bakalářská práce. Vysoké učení technické v Brně, Fakulta chemická, Ústav fyzikální a spotřební chemie. Vedoucí práce Václav Brázda.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a že všechny použité zdroje jsem správně a úplně citoval. Bakalářská práce je z hlediska obsahu majetkem Fakulty chemické VUT v Brně a může být využita ke komerčním účelům jen se souhlasem vedoucího bakalářské práce a děkana FCH VUT.

.....

Podpis studenta

Poděkování: Děkuji doc. Mgr. Václavu Brázdovi, Ph.D. za odborné vedení, cenné rady a poskytnutí velice zajímavého tématu pro bakalářskou práci. Velké díky také patří rodině a přátelům, kteří mi s prací, ať už v jakékoliv míře, pomohli. Zvláštní poděkování náleží Filipu Richterovi za jeho výpomoc při zpracování dat a Michalu Procházkovi za rady týkající se problematik molekulární biologie.

Obsah

1. Úvod

Cupriavidus Necator H16, starším názvem *Ralstonia Eutropha* H16 (dále jen CN H16), je gram-negativní půdní bakterie známá jako producent polyhydroxyalkanoátů (dále jen PHA). Tyto polymerní zásobní látky mají v průmyslu velký potenciál jako materiály pro syntézu degradabilních bioplastů šetrných k životnímu prostředí. CN-H16 je schopný za vhodných podmínek syntetizovat polyhydroxyalkanoáty až do 90% suché váhy buňky [1]. V zájmu průmyslového využití byl pro účely genového inženýrství sekvenován a publikován genom CN-H16 už několika různými laboratořemi [2][3]. Přítomnost těchto veřejně dostupných referenčních genomů značně ulehčuje analýzu genomu CN-H16 pro případné změny.

Jedna z metod použitelná pro tento účel je „Next Generation Sequencing“ (dále jen NGS). Jejím výsledkem je stanovení celé genomové sekvence daného organismu. Porovnáním takto stanovených sekvencí se dají určit mutační změny odlišných populací. Na základě vhodných referencí z jiných experimentů a velkého množství sekvenované informace lze vyřadit případné artefakty sekvenování a vyhodnotit věrohodnost dat [4].

Cílem této práce je analyzovat data z NGS tří různých populací CN H16 pocházejících z experimentu Ing. Nováčkové a docenta Obruči [5]. Jde o „wild-type“ populaci pocházející ze začátku experimentu (CN-wt), populaci adaptovanou na osmotický stres (CN-Na41) a populaci adaptovanou na stres vyvolaný těžkými kovy (CN-Cu44). Dalšími cíli této práce jsou stanovení genetických změn adaptovaných populací CN-Na41 a CN-Cu44 oproti původnímu genomu CN-wt a stanovení odlišností genomu CN-wt oproti referenčnímu genomu z databáze NCBI [3].

2. Teoretická část

2.1. Bakteriální genom a jeho složení

Bakteriální genom je soubor veškeré genetické informace, kterou má bakteriální buňka k dispozici. Genetická informace je uložena v lineárních/cirkulárních molekulách deoxyribonukleové kyseliny (dále jen DNA) zvaných chromozomy, případně v DNA plasmidech. Nejobvyklejší struktura genomu v bakteriálních buňkách je kruhový chromozom [6]. Jsou však i druhy bakterií s chromozomy lineární struktury s repetitivními (telomerickými) sekvencemi na koncích [7]. Plasmidy jsou malé molekuly DNA oddělené od chromozomů, které se nezávisle replikují a obvykle obsahují genetickou informaci pro tvorbu jen několika málo proteinů, například zvyšujících rezistenci k antibiotikům [8].

Chromozomy i plasmidy bakterií jsou tvořeny dvoušroubovicovou molekulou DNA. DNA obsahuje nukleotidové báze (A,C,G a T), jejichž sekvence podle pravidel genetického kódu poskytují dané bakterii informaci o struktuře proteinů a funkčních typů RNA, nebo případně fungují jako nekódující regulační úseky [9]. Bakteriální genomy obecně obsahují zhruba 10^5 až 10^7 párů bází. Jsou tvořené z většiny kódujícími sekvencemi a neobsahují, na rozdíl od například eukaryotických buněk mnohobuněčných organismů, mnoho nekódujících opakujících se sekvencí [10]. Částečně kvůli těmto vlastnostem bakteriálních genomů a díky pokrokům v sekvenačních technologiích poslední doby se sekvenování celých bakteriálních genomů stalo obvyklou a cenově dostupnou laboratorní metodou [11].

2.2. Základní typy mutací v bakteriálních genomech

Mutace v bakteriálních genomech jsou obecně změny v genetické informaci nezpůsobené transposony, integrony nebo rekombinací. Dělí se na mutace substituční, inzerční a deleční. Substituční mutace je taková mutace, při které dojde k záměně jednoho nukleotidu za jiný. Toto se může stát například při tautomerizaci dusíkaté báze a následného přiřazení nukleotidu při replikaci, který za normálních podmínek není komplementární. Při bodové mutaci tedy nedochází k posunu sekvence, na rozdíl od inzerčních mutací, kdy se dovnitř sekvence vkládají dodatečné nukleotidy, nebo delečních mutací, kdy se zevnitř sekvence nukleotidy odebírají. Inzerční a deleční mutace malých rozměrů (1-50 párů bází) vznikají například kvůli

“sklouzávání“ DNA polymerázy v oblastech s opakujícími se nukleotidy, kdy komplementarita bází neudrží vlákna ve správné pozici vůči sobě [12].

Mutace vznikají buďto vlivem určitých chemikálií či záření, které ovlivňují chemickou strukturu DNA nebo jako chyby při replikaci, které nejsou opraveny. Replikace v bakteriích s kruhovými chromozomy probíhá z jediného místa a v obou směrech, tedy za vzniku dvou replikačních vidlic. V replikační vidlici se dělí dvoušroubovice DNA na vedoucí a opožďující vlákno, protože bakteriální DNA polymeráza syntetizuje komplementární vlákno pouze ve směru 5'→3'. Vlákno syntetizované ve směru 3'→5' tedy musí být syntetizováno po částech zvaných Okazakiho fragmenty. Tyto fragmenty jsou poté spojovány DNA ligázou. Studie s bakterií *E. coli* ukázaly, že samovolné mutace (nevyvolané externími mutageny) vznikají na opožďujícím vlákně přibližně 20krát častěji [12][13].

2.3. Význam rekombinantních bakteriálních mutantů v průmyslu

Cílená evoluce bakterií skrze mutace a rekombinace je v mikrobiálním průmyslu klíčová metoda. Obvyklý postup je chemické vyvolávání mutací v určité mikrobiální populaci s následnou selekcí nejvhodnějšího kmene s nejvyšší produkcí žádané látky. Postup cílené mutageneze a selekce může zvýšit produkci žádané látky až o několik řádů [14].

Poslední dobou se tento přístup kombinuje s moderními molekulárně-biologickými metodami rekombinace. Příkladem je elektroporace bakterií – metoda, která pomocí silného elektrického pulzu krátkodobě vytvoří póry v membráně bakteriální buňky a tím umožní vstup cizorodé DNA, jako jsou například plasmidy. Tento způsob transformace je mnohem efektivnější než chemická transformace [15][16]. Typické využití rekombinantních bakterií je na produkci proteinů, které se obtížně získávají přírodními metodami. Příkladem může být bakterie *Escherichia coli*, do které roku 1978 Herbert Boyer úspěšně vložil lidský transgen kódující inzulin. Produkce “syntetického“ lidského inzulinu tímto způsobem byla v USA povolena o čtyři roky později [17][18].

Na využití geneticky modifikovaných bakterií v průmyslu se vztahují různě přísné zákony. Například v USA musí být geneticky modifikované mikroby schváleny vládní agenturou Environmental Protection Agency. Ačkoliv některé země průmyslové využití geneticky modifikovaných mikroorganismů úplně zakazují, jejich výzkum v laboratorním prostředí zakázaný nebývá [19].

2.4. Historie genomového sekvenování

Roku 1953, krátce po objevu struktury DNA, George Gamow odhalil způsob přepisu genetické informace z nukleových kyselin do proteinů objevem tripletového genetického kódu. S tímto objevem se zvýšil zájem o metodu, kterou by se genomové sekvence daly stanovovat. První experimenty v této oblasti se zaměřovaly na sekvenování RNA (především kvůli dostupnosti RNáz a nepřítomnosti komplementárního vlákna). V roce 1965 Robert Holley et al. publikovali první celou kódující sekvenci nukleotidů náležící tRNA alaninu z kvasinky *Saccharomyces Cerevisiae*. Walter Fiers et al. v roce 1976 publikovali první kompletní DNA genom, náležící bakteriofágu MS2 [20][21].

Nejdůležitější objev této doby ovšem přišel později v roce 1977, kdy Frederick Sanger přišel s novou metodou založenou na selektivním přidávání radioaktivně značených dideoxynukleotidů DNA polymerázou k fragmentům stanovované DNA sekvence. Směs se po přidání těchto značených dideoxynukleotidů separuje elektroforézou podle velikosti. Následně se podle radioaktivního značení koncového dideoxynukleotidu, který brání další elongaci fragmentu, určí poslední báze každého fragmentu. Za přítomnosti dostatečného množství fragmentů různé délky bylo tímto způsobem možné jednoduše a rychle stanovit pořadí až několika set bází v nukleových kyselinách. Sanger et al. krátce po tomto objevu stanovili sekvenci lidské mitochondriální DNA (16 569 párů bází) a DNA bakteriofágu lambda (48 502 párů bází). Tato metoda se nazývá Sangerovo sekvenování a byla mimo jiné použita pro první sekvenaci lidského genomu v rámci „The Human Genome Project“ mezi lety 1990 a 2003. Díky možnosti automatizace a s dalšími modifikacemi je Sangerovo sekvenování základem i mnoha současných NGS metod [20][22].

Během dalších let byly vyvíjeny nové metody upravující a zdokonalující postup sekvenování na základě selektivního přidávání nukleotidů – také známé jako sekvenování syntézou. Milníkem byla metoda pyrosekvenování, poprvé popsaná v roce 1993 a později licencovaná firmou 454 Life Sciences. Tato metoda je založena na sekvenování syntézou a detekci vznikajícího pyrofosfátu. Pyrofosfát je převáděn na ATP, které následně reaguje s luciferázou a kyslíkem za vyzařování fotonů. Takto vzniklé fotony jsou detekovány například fotonásobičem. Na základě nukleotidů přidaných v daném kroku a intenzitě světelného toku dopadajícího na detektor je stanovena sekvence DNA. Pyrosekvenování se stalo první komerčně úspěšnou metodou sekvenování genomů a způsobilo značný pokles ceny sekvenování. Po pozdějších objevech se nakonec v roce 2013 od metody upustilo, protože

přestala být konkurenceschopná [23]. Po úvodním úspěchu metody pyrosekvenování vznikly totiž metody další a levnější, například metoda „Solexa sequencing“, později licencovaná společností Illumina [20].

Metody dnešní doby, jako například „Illumina Short-Read“ nebo „Ion Torrent Sequencing“, se vyznačují především schopností masivně paralelního sekvenování krátkých fragmentů. Při těchto metodách je obrovské množství krátkých fragmentů (zhruba 50 až 100 párů bází) stanovované molekuly DNA přichyceno v mikrocele pomocí specifických amplifikačních primerů a sekvenováno multiparalelně. Mezi moderní metody odlišného typu patří například „Single Molecule Real Time Sequencing“ od firmy Pacific Bio (obecně známo jako “PacBio Sequencing“), při které se sekvenuje jeden fragment DNA dlouhý až 10 000 párů bází. Další metodou je například tzv. „SOLiD sequencing“ od firmy Life Technologies, která je založená na schopnosti DNA ligázy rozpoznat chybné báze na komplementárním vlákně a probíhá bez syntézy komplementárního vlákna. Díky moderním metodám (obecně zvaným NGS), se časové a cenové nároky na sekvenování drasticky snížily. Lidský genom je dnes možné celý sekvenovat za cenu nižší než 1000 amerických dolarů s odběrem vzorku a zpracováním dat během několika dnů [20][24].

2.5. NGS metody Illumina short-read

„Read“ je klíčový pojem v sekvenačních metodách, jelikož jde o počítačový záznam popisující určitý osekvenovaný DNA fragment. Soubor všech readů z jedné sekvenace je kompletní výstup z NGS, a proto jakékoliv hledané informace o sekvenovaném genomu se získávají zpracováním readů [25]. Délka a relativní četnost readů vůči velikosti sekvenovaného genomu jsou jedny z hlavních rysů dané metody NGS, jelikož do značné míry určují, jak výstupní data budou zpracována. U metod s krátkými ready (tzv. „short-read“ metody) může být problém se sestavením genomu, pokud se v genomu vyskytují repetice delší než ready [26]. Častý způsob minimalizace tohoto problému je sekvenování krátkých readů z konců jednoho dlouhého fragmentu. Například namísto úplného sekvenování fragmentů dlouhých 200 bází se sekvenuje 100 bází z každého konce jednoho fragmentu dlouhého 600 bází. Výsledkem tedy jsou dva ready dlouhé 100 bází s klíčovou informací, že mezi nimi musí být v sestaveném genomu 400 bází mezera (tzv. „Paired-End“ ready), což je pro rozluštění oblastí s repeticemi velice užitečné [25]. Obecně výhodami sekvenačních metod s krátkými ready jsou nízká cena a vysoká rychlost [27].

Short-read metody firmy Illumina jsou založené na sekvenování syntézou neboli SBS („sequencing by synthesis“) [28]. Při experimentu se vzorek nejprve rozštěpí na náhodných místech a připojí se adaptérové sekvence, které následně komplementární vazbou umožní připevnění na povrch průtokové cely, ve které SBS proběhne. Fragmenty bez adaptérových sekvencí se na povrch nepřipojí a jsou z průtokové cely promýváním při SBS odstraněny. Následně proběhne tvorba klastrů (skupin) skrze „bridge PCR“ amplifikaci. Bridge PCR spočívá v tom, že každý připevněný fragment DNA v průtokové cele má adaptérové sekvence na obou koncích a povrch průtokové cely má spoustu oligonukleotidů schopných adaptérové sekvence vázat. Po každé syntéze komplementárního řetězce se následnou denaturací vytvoří dvojnásobek ssDNA fragmentů, v ideálním případě výhradně identických či komplementárních k původnímu fragmentu. Tímto způsobem vznikají z jednotlivých fragmentů tisíce klastrů upevněných na průtokovou celu. Následně proběhne samotné sekvenování syntézou za pomoci reverzibilních terminátorů – deoxynukleotidů s blokovaným 3‘ OH koncem. Průtokovou celou jsou vždy promývány všechny čtyři možné reverzibilní terminátory, které jsou vázané DNA polymerázami na nové řetězce podle komplementárních vláken. Následně je průtoková cela osvícena vhodnými vlnovými délkami pro excitaci každého ze čtyř fluorescentně značených reverzibilních terminátorů. Intenzita fluorescence pro každý klastr je zaznamenána detektorem a z reverzibilního terminátoru je následně odstraněna fluorescenční (a elongaci zabraňující) část. Opakování tohoto procesu tedy vede k syntéze celého komplementárního vlákna pro každý fragment. Následně počítač vyhodnocuje fluorescenční stopy a přiřazuje nejpravděpodobnější bázi a skóre kvality vyjadřující jistotu určení báze [28].

Výsledkem metody jsou tedy ready obsahující sekvence bází A, C, G, T (nebo N v případě neznámé báze) se zaznamenanou kvalitou určení báze. Nejistoty mohou vznikat při neúplné vázbě komplementárních nukleotidů, snížením enzymatické aktivity ke konci syntézy komplementárního řetězce nebo systematickým či náhodným selháním instrumentace, jako například vznikem bubliny v průtokové cele [27].

2.6. Výstupní data z NGS

Základním datovým formátem sekvenační dat je FASTQ, který obsahuje informace o snímaných readech. Koncovka souborů tohoto formátu je .fastq a jelikož pro každý read jsou v souboru uvedeny mimo jiné všechny báze daného readu, kvalita každé báze (číslo popisující pravděpodobnost správného určení báze), číslo označující odkud read pochází z průtokové cely,

případně zda read náleží do páru, velikost těchto souborů je v řádu gigabajtů. FASTQ soubory jsou obvykle "očištěny" od nekvalitních readů a následně použity pro porovnání readů z konkrétního místa vůči stejnému místu na vhodném referenčním genomu (mapování) za účelem kvantifikování mutací nebo pro „assembly" [29].

Assembly je proces jakým algoritmy z readů sestavují původní genomovou sekvenci na základě jejich překryvů. Sekvence takto získané však nebývají totožné s genomem, z jakého ready pochází, ale naopak rozdělené na desítky částí. Tyto části se nazývají „scaffoldy" a algoritmy nejsou schopné je korektně složit na původní genom primárně kvůli ztrátě informace spojené s fragmentováním genomu v chemické části NGS [26]. Pro sestavení sekvence totožné s původním genomem je potřeba NGS zkombinovat s dalšími molekulárně-biologickými metodami nebo využít vhodné reference z jiných experimentů (strukturně podobné genomy) [26][30]. Formát v jakém assembly ukládá výsledky se nazývá FASTA a jeho odpovídající koncovky jsou například .fasta, .fa nebo .fna. FASTA je formát, který obsahuje pouze sekvenci nukleotidů. Vyskytují se v něm tedy písmena A, C, G, T nebo N a ukládají se v něm genomy, exomy, transkriptomy či jejich části [31].

Dalšími formáty získanými zpracováním readů jsou BAM a SAM, neboli Binary Alignment Map a Sequence Alignment Map s odpovídajícími koncovkami .sam a .bam. Oba tyto formáty splňují stejný účel, což je mapování readů k nějaké vhodné referenční sekvenci. Mapování je proces, při kterém se algoritmus snaží pro každý read na referenční sekvenci najít místo kam patří. Každému takto přiřazenému readu se dále říká „alignment" a program mu přiřadí kvalitu mapování, která určuje jeho správnost. Tato kvalita záleží například na shodě mezi bázemi readu a bázemi daného místa na referenční sekvenci. Výstupní formáty mapování se liší se pouze v kódování, BAM je kódovaný úsporněji a má zhruba o řád menší velikost než SAM, který má postradatelnou výhodu, že je čitelný člověkem. Při vytvoření souboru tohoto formátu z readů nedochází ke ztrátě informace jako při assembly. Formát SAM je organizován tak, že v každém řádku je uvedena jedna pozice v referenčním genomu, odpovídající referenční báze na té pozici a dále informace získané z readů odpovídající dané pozici. Obecně se tento formát používá ke stanovování mutací pomocí algoritmů, které s nastavenou citlivostí soubor prohledají a vypíší, kde se báze z readů neshodují s referenční bází [29].

Pro tento postup hledání mutací („Variant Calling") byl vytvořen formát VCF („Variant Calling Format") s koncovkou .vcf. Jde o jednoduchý formát s hledanými řádky ze SAM souboru, výrazně upravenými tak, aby byly pro člověka lépe čitelné. Tento formát pro

zjednodušení stanovování mutací obsahuje i další dopočítané informace, například frekvenci každé mutace a průměrnou kvalitu bází z readů [32].

Ačkoliv účely jednotlivých formátů zůstávají napříč bioinformatickou sférou stejné, konkrétní způsoby kódování a obsažené informace se mohou lišit, zejména v závislosti na podmínkách použité metody NGS [29]. Veřejně dostupné bioinformatické nástroje jsou obecně na odlišná formátování připravené, ale vzhledem k tomu, že ne všechny formáty jsou jednoznačně definované, tak někdy může někdy k problémům s kompatibilitou.

2.7. Bioinformatické nástroje pro zpracování dat z NGS

V dnešní době jsou veřejně dostupné bioinformatické nástroje pro jakýkoliv účel bioinformatické analýzy společně s extenzivní, detailní dokumentací, návody a recenzemi. Ačkoliv jsou tyto návody obvykle určeny pro biology, často jejich autoři předpokládají určitou znalost informatiky. Příkladem může být článek: "Beginner's guide to comparative bacterial genome analysis using next-generation sequence data" [33], kde autoři v manuálu doprovázejícím článek vysvětlují jednotlivé kroky bakteriální genomové analýzy někdy až s vyčerpávajícími detaily. Přestože z prvního čtení je očividné, že článek je určen pro úplné začátečníky informatiky, tak už na čtvrté straně, v doporučeném postupu assembly pomocí programu Velvet, autoři vynechali klíčovou informaci, že program musí být zkompilován za určitých (ve článku nezmíněných) parametrů, což není triviální záležitost na pochopení a provedení pro většinu začátečníků v informatice [34].

V kontextu problematiky zpřístupnění bioinformatických nástrojů výzkumníkům stojí za zmínku webový portál Galaxy [35]. Jde o projekt tří univerzit z USA (Penn State University, John Hopkins University, Oregon Health & Science University) s cílem sdružit užitečné a veřejně dostupné nástroje a v přehledné formě je zpřístupnit na jednom internetovém portálu. Používání programů na Galaxy je ve většině případů pro uživatele výrazně zjednodušeno, například automatickým doplňováním různých méně podstatných parametrů. Výše zmiňovaná nutnost kompilace pro různé programy a podobné "informatické formality" jsou dělány automaticky. Další výhodou je uchovávání dat na online serverech a tím pádem nepotřebnost zatěžování vlastního počítače výpočetně či pamětně. Galaxy také doporučuje programy podle jejich návaznosti, například pokud uživatel použije program na statistickou analýzu určitých dat, Galaxy doporučí použití programu na vizualizaci statistické analýzy. Nevýhodou Galaxy

může být poněkud velký počet dostupných nástrojů stejného typu a příliš stručné popisky jejich funkcí. Uživatel tedy musí mít pro pohodlné použití Galaxy alespoň minimální orientaci v bioinformatických nástrojích které používá [36].

Ukázkový postup vyžadující bioinformatické nástroje je analýza mutací z jedné nebo několika sekvenovaných populací porovnáním s vhodným referenčním genomem. Pro tuto analýzu je tedy potřeba FASTA sekvence vhodného referenčního genomu a FASTQ soubory pro každou populaci, která má být vůči referenční sekvenci porovnávána. Prvním krokem v této analýze je kontrola kvality readů, například pomocí programu FASTQC [37]. Po zkontrolování kvality a případném upravení FASTQ souborů například pomocí programu Fastp [38], je na místě vytvoření BAM souborů a jejich zkombinování na soubor typu „pile-up“. Obojí se dá zvládnout například pomocí SAMTools [39]. Prohledáním pile-up souboru pomocí například programu VarScan 2 [40], jehož výstup je ve formátu VCF, je analýza zhotovena. Všechny zmíněné nástroje a programy jsou dostupné na Galaxy. Podobné analýzy se samozřejmě dají provést pomocí jiných nástrojů nebo jinými postupy. Pro jistotu správnosti výsledků může být vhodné porovnat výsledky získané různými postupy [41].

2.8. Produkce polyhydroxyalkanoátů u CN H16

Bakteriální buňky disponují genetickými komponentami vysoce specializovanými na tvorbu různých produktů s přidanou hodnotou a velkým významem v biotechnologickém průmyslu [42]. Jedním z těchto produktů jsou polyhydroxyalkanoáty (PHA) a bakterie s genetickou výbavou nutnou pro jejich syntézu se v posledních desetiletích s pokroky v metodách molekulární biologie opět stávají cílem mnoha experimentů. Tyto experimenty jsou motivované získáním vhodnějších genotypů pro průmyslovou syntézu PHA pomocí metod genového inženýrství [43].

Polyhydroxyalkanoáty jsou biodegradabilní polyestery syntetizované bakteriemi při stresových podmínkách jako zásoby uhlíku a energie [5]. V bakteriích se tyto zásoby PHA nacházejí ve formě granulí a plní nejen zásobní funkce, například zvyšují odolnost vůči osmotickému stresu [44]. PHA jsou primárně tvořeny 3-hydroxyacyl monomery krátkého (C_3 až C_5), středního (C_6 až C_{12}) nebo dlouhého uhlíkového řetězce (C_{13} a více), které jsou spojovány v různých poměrech enzymem PHA syntázou do ko-polymerních jednotek za vzniku granulí. Ačkoliv ve výsledku délky řetězců monomerů záleží na struktuře dostupného substrátu,

nejčastějším monomerem je obecně 3-hydroxybutyrát (3HB) s methylovou skupinou na místě vedlejšího řetězce [43]. Poly-3-hydroxybutyrát má sám o sobě výhody oproti různým petrochemicky získaným polymerům (biodegradabilita například), je však velice křehký. Tento problém se dá řešit například inkorporací 3-hydroxyvalerátu přidáním propionové či valerové kyseliny do růstového média za vzniku kopolymeru poly-hydroxybutyrát-hydroxyvalerát (PHBV) [45].

CN H16 je gram-negativní, chemolitotrofní bakterie a obecně jeden z nejpoužívanějších modelových organismusmů pro studium produkce polyhydroxyalkanoátů [43] kvůli jeho ojedinělým výhodám oproti ostatním producentům. CN H16, na rozdíl od ostatních bakterií z rodu *Cupriavidus*, například *C. pauculus*, *C. gilardii* a *C. metallidurans* není patogenní pro člověka. Další jeho výhodou je, že postrádá genetickou výbavu pro syntézu alternativních látek (jiných než PHA) z 3-hydroxyacyl-CoA za stresových podmínek, jako můžou být u jiných bakterií například rhamnolipidy nebo alginát [42]. CN H16 tedy tvoří granule PHA bez těchto kontaminantů a to (za extrémních podmínek) až do 90% suché váhy buňky [1][46]. Další výhoda je, že genom CN H16 je velice dobře zmapovaný. Nově anotovaný genom CN H16 z roku 2019 je dostupný pro veřejnost na stránkách NCBI [3].

2.9. Geny související s akumulací PHA v CN H16

Geny zodpovědné za většinu tvorby PHA v CN H16, ve formě poly-3-hydroxybutyrátu (PHB), leží v konstitutivním (stále aktivním) phaCAB operonu [42]. Dle pořadí, v jakém příslušné enzymy působí v metabolické dráze, tyto geny kódují:

- acetyl-CoA acetyltransferázu (ketothiolázu), která katalyzuje kondenzaci acetyl-CoA na acetoacetyl-CoA, příslušný gen je phaA
- acetoacetyl-CoA reduktázu, která (za vstupu NADPH a výstupu $NADP^+$) katalyzuje redukci acetoacetyl-CoA na 3-hydroxybutyryl-CoA, příslušný gen je phaB
- PHA syntázu, která monomer 3-hydroxybutyryl-CoA v počtech více jak 15-ti [42] spojuje na polymerní jednotky poly-3-hydroxy-burátu, příslušný gen je phaC

Geny ale nejsou v operonu v tomto pořadí, nýbrž v pořadí phaC→ phaA→ phaB, viz Tabulka 1.

Zkratka genu	Pozice první báze	Pozice poslední báze	Délka	Kódovaný protein
phaP1	1 498 220	1 497 642	579	phasin P1
phaC	1 556 003	1 557 772	1770	PHA syntáza
phaA	1 557 857	1 559 038	1182	acetyl-CoA acetyl-transferáza
phaB	1 559 113	1 559 853	741	acetoacetyl-CoA reduktáza
phaR	1 560 290	1 560 841	552	transkripční faktor

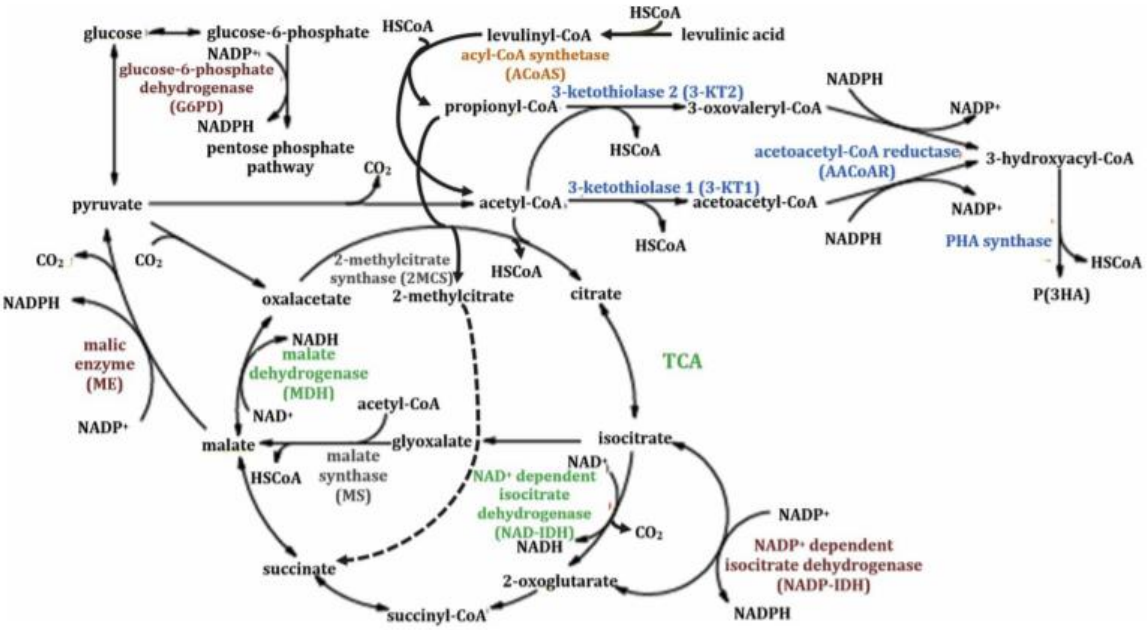
Tabulka 1 - pozice genů operonu phaCAB, phaP1 a phaR na chromozomu 1 v genomu CN H16 z NCBI [3]

Dále hned za operonem se nachází gen phaR, který kóduje regulační protein tlumící transkripci phaR a phaP1. PhaP1 gen má čtyři homology v genomu CN H16 a ty kódují „phasin" proteiny, které se vážou na povrch granulí PHA, stabilizují je a brání koalescenci. Tato funkce je důležitá pro akumulaci PHA, jelikož větší počet granulí má pozitivní efekt na aktivitu PHA syntázy [43]. Regulace phaP1 probíhá skrze protein kódovaný genem phaR, který je také schopný podobné funkce jako phasin proteiny a při tvorbě nových granulí PHA se sníží jeho koncentrace v cytoplazmě (inverzně se zvýší na granulích PHA) a dojde k zvýšené syntéze phaP1 a phaR příslušných proteinů. Při saturaci povrchů granulí se nakonec koncentrace phaR proteinu v cytoplazmě opět zvýší a dojde k tlumení transkripce genů phaR a phaP1 [47][48]. Zvýšená exprese genu phaP1 vede k akumulaci spousty malých granulí a mutace poškozující funkci tohoto genu vede k akumulaci pouze jedné velké granule PHA [43]. Populace CN H16 s funkčně podobnou mutací rychleji degradují PHA a akumulují jej méně [49].

Genom CN H16 také obsahuje gen phaZ kodující PHA depolymerázu, jejíž funkce je odebírat ze zásobních granulí jednotlivé monomery na zpracování. Tato depolymeráza má několik homologů, které se dělí na intracelulární a extracelulární a mají klíčovou funkci při degradaci PHA v přírodě [50].

Mimo jiné produkce PHA může být ovlivněna redoxní rovnováhou v buňce nebo pochody citrátového cyklu. Například mutace oslabující funkci isocitrát dehydrogenázy v CN H16 způsobují nahromadění acetyl-CoA a jeho zvýšený tok do metabolické dráhy syntézy PHA [43], viz Obrázek 1 - schéma metabolismu PHA u CN H16 s vybranými enzymy, zdroj: [5]. CN H16 je schopný za dostupnosti vhodných substrátů syntetizovat i jiné PHA monomery než poly-3-hydroxybutyrát (například 3-hydroxyvalerát, 4-hydroxybutyrát) pomocí odpovídajících ketothioláz. Mutanti s neschopností využití 3-hydroxyvalerátu nebo 4-hydroxybutyrátu jako

zdroje uhlíku inkorporují mnohem větší množství těchto látek do granulí PHA [43]. Tento fakt má potenciální využití v kontrolované průmyslové výrobě PHA specifického složení.



Obrázek 1 - schéma metabolismu PHA u CN H16 s vybranými enzymy, zdroj: [5]

3. Cíl práce

Cílem této práce je analýza mutačních změn ve dvou adaptovaných populacích bakterie *Cupriavidus necator* H16 (CN-Cu44 a CN-Na41) v porovnání s jejich původním kmenem (CN-wt). Stanovení těchto mutačních změn je provedeno bioinformatickou analýzou dat z NGS těchto tří populací.
Dílčí úkoly zpracované v rámci této práce se dají shrnout následovně:

- vyhodnocení kvality analyzovaných NGS dat
- stanovení konkrétních mutačních změn adaptovaných populací oproti původní populaci
- asociování mutačních změn s konkrétními oblastmi na genomu CN H16 a jejich funkcemi
- diskuse případných dopadů stanovených mutačních změn na produkci PHA

4. Praktická část

4.1. Zdroj NGS dat - experiment Ing. Nováčkové

Experiment, z něhož pochází data analyzovaná v praktické části této práce, byl zaměřen na evoluční inženýrství bakterie CN H16. Experiment byl prováděn za účelem zkoumání evoluční adaptace CN H16 za specifických stresových podmínek a s účelem případného získání adaptovaného kmenu s lepšími schopnostmi akumulace PHA či vyšší odolností proti vybraným stresovým podmínkám. Jako stresové faktory byla přitom využita měď v podobě Cu^{2+} představující antropogenní polutant a chlorid sodný (NaCl) reprezentující přirozeně se vyskytující stresový faktor v kontextu průmyslových fermentací.

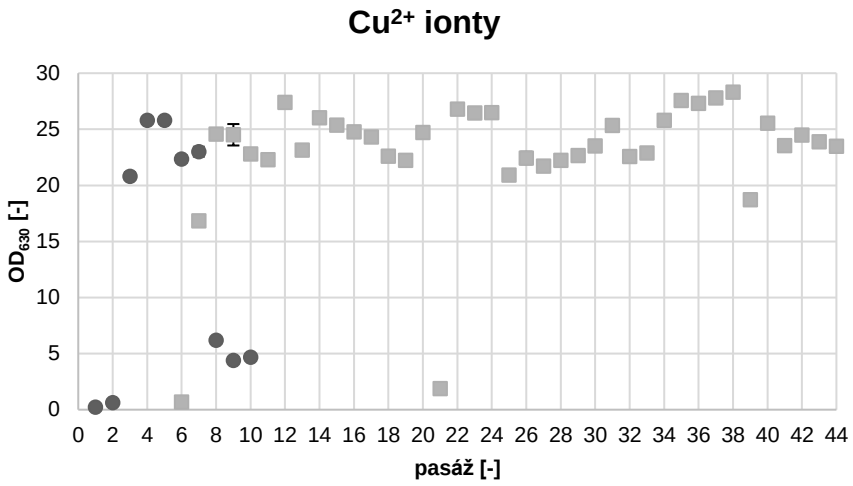
4.1.1. Postup experimentu a výsledky

Při samotném experimentu byl sbírkový kmen CN H16 (CCM 3726) kultivován submerzním způsobem ve dvou liniích za přítomnosti specifických stresorů, a to následujícím způsobem:

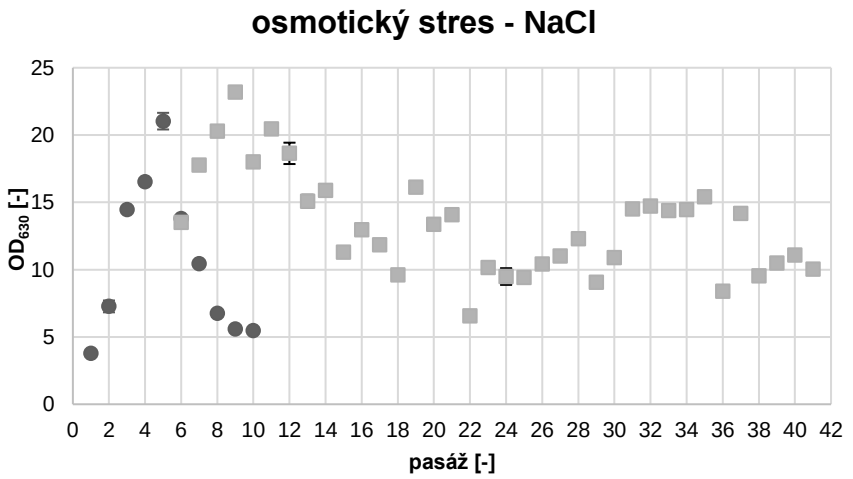
- První z linií byla kultivována v médiu s osmotickým stresem vyvolaným NaCl koncentrace 20 g/L.
- Druhá z linií byla kultivována v médiu se stresem vyvolaným Cu^{2+} ionty koncentrace 30 mg/L.

V obou případech byly populace přeočkovány po každých 48 hodinách růstu za zachování konstantních koncentrací stresorů. Jako zdroj uhlíku byla použita fruktóza o koncentraci 20 g/L. Tyto stresové podmínky byly mimo jiné vybrány, protože osmotický stres i Cu^{2+} ionty byly prokázány jako podmínky ovlivňující akumulaci PHA u různých bakterií [51] [52]. Selekční tlak byl zachován při experimentu konstantní, jelikož při začátku došlo zvyšováním selekčního tlaku k odumírání populací. U páté až sedmé pasáže došlo k odumření, a proto byly rozočkovány páté pasáže z kryozkumavek při jejich původních koncentracích stresorů (uvedených výše).

Pro porovnání růstu adaptujících linií byla měřena optická hustota při 630 nm (OD630) jednotlivých pasáží:

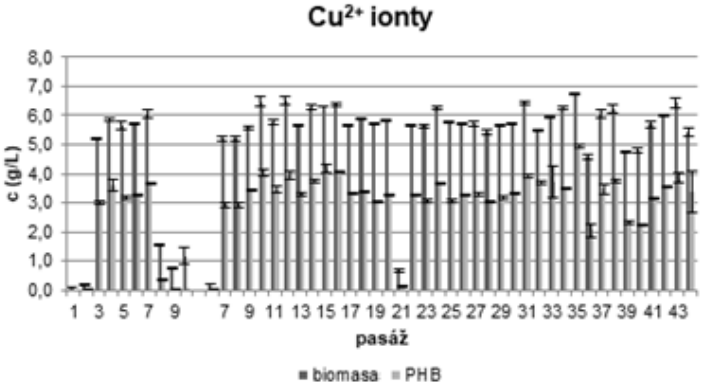


Obrázek 2 - Graf závislosti OD630 na pasáži adaptace na Cu2+ stres, kolečka značí populace adaptující na zvyšující selekční tlak, čtverce značí populace adaptující na konstantní selekční tlak

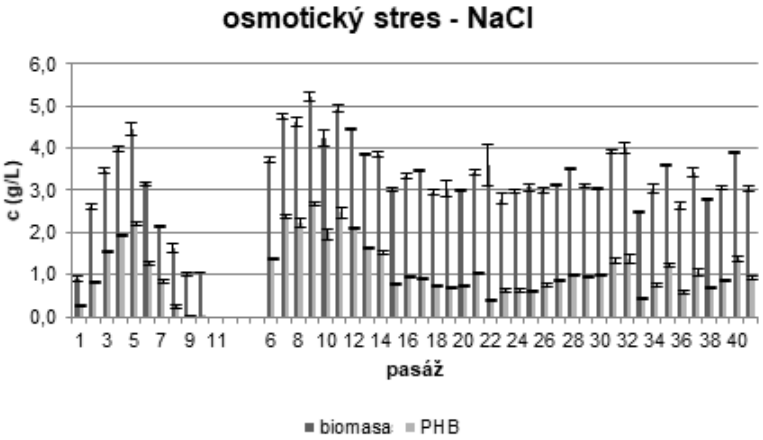


Obrázek 3 - Graf závislosti OD630 na pasáži adaptace na osmotický stres kolečka značí populace adaptující na zvyšující selekční tlak, čtverce značí populace adaptující na konstantní selekční tlak

Dále byl gravimetricky stanoven obsah biomasy v jednotlivých pasážích:



Obrázek 4 - Graf závislosti koncentrace biomasy a PHB na pasáži adaptace na Cu2+ stres



Obrázek 5 - Graf závislosti koncentrace biomasy a PHB na pasáži adaptace na osmotický stres

Obrázky 2 až 5 byly přeloženy do češtiny, přičemž jejich autorem je Ing. Nováčková.

4.1.2. Populace odeslané na NGS

Pro vyhodnocení genetických změn v adaptovaných pasážích byly ze tří populací vyizolovány vzorky DNA a odeslány na NGS. Tyto populace jsou:

- původní populace ze začátku experimentu (CN-wt)
- 44. pasáž adaptace na Cu^{2+} stres (CN-Cu44)
- 41. pasáž adaptace na NaCl stres (CN-Na41)

Pro účely izolace byla příslušná kultura zaočkována z kryozkumavky do NB („Nutrient Broth“) média, kde byla kultivována přibližně 24 hodin při 180 rpm. Následně bylo inokulum přeočkováno do minerálního média, kde bylo kultivováno 48 hodin opět při 180 rpm. Poté byly do sterilní eppendorfky centrifugovány 2 mL kultury při 6000 rpm po dobu 5 minut a výsledný pelet byl použit pro izolaci genomové DNA pomocí kitu Macherey-Nagel™ NucleoSpin™ Microbial DNA. Postup tohoto kitu je založený na mechanické lýze buněk skleněnými kuličkami při centrifugaci a izolaci DNA skrze adsorpci na křemičité sklo. Vyizolovaná DNA byla analyzována elektroforézou a nanodropem pro vyhodnocení kvality a kvantity DNA ve vzorku.

V konečné fázi byly všechny vzorky připraveny na NGS dle požadavků firmy a odeslány k analýze. Doba kultivování adaptovaných populací při experimentu koreluje se zhruba tisícem generací. Výsledkem NGS byly data zpracovávaná v následujících kapitolách.

4.2. Popis analyzovaných dat z NGS

Jako součást NGS analýzy našich tří populací společnost DNALink [53] (kromě provedení samotného NGS) zpracovala NGS data pomocí různých bioinformatických metod. Všechny tři bakteriální populace byly zpracovány identickým postupem, který soudě dle struktury výstupních dat a informací podaných společností DNALink samotnou, byl zhruba následující:

1. Mechanická fragmentace vzorku DNA na fragmenty průměrné délky 600 bází a ligace adaptérových sekvencí pomocí kitu Truseq Nano DNA Prep Kit,
2. NGS – provedeno na Illumina sekvenátoru Novaseq 6000 metodou párových readů délky 101,
3. de-novo assembly genomové FASTA sekvence z NGS dat pomocí programu Unicycler,
4. predikce genů z pořadí nukleotidů ve FASTA sekvenci,
5. popis (anotace) predikovaných genů na základě podobnosti se známými geny z databáze (pravděpodobně NCBI databáze [64], případně speciální databáze společnosti),
6. hledání mutací (variant calling) v populaci na základě porovnávání readů z NGS vůči sestavené FASTA sekvenci.

Výsledkem těchto metod byla jedna složka s daty a informacemi pro každou populaci. Každá tato složka obsahuje následující soubory:

- složku „Assembly“ s FASTA sekvencí genomu získanou pomocí de-novo assembly a s FASTA sekvencemi jednotlivých scaffoldů (přibližně 50),
- složku „rawdata“ s dvěma soubory FASTQ z Illumina short-read sekvenace, označenými R1 a R2 obsahujícími párové ready z každého konce snímaného fragmentu,
- složku „GenePrediction“ s upravenými FASTA sekvencemi scaffoldů rozdělenými na očíslované predikované geny (otevřené čtecí rámce),
- složku „GeneAnnotation“ s textovými a excelovými soubory obsahujícími informace o jednotlivých čtecích rámcích,
- složku „VariantDiscovery“ obsahující VCF soubor detailně popisující jednotlivé mutace stanovené z readů a FASTA sekvence daného genomu a excelovou tabulkou shrnující některé informace z VCF souboru,
- PDF soubor „Analysis_report“ obsahující informace o způsobu provedení metody NGS, o readech a všech provedených (výše zmíněných) metodách.

Od společnosti nám tedy byla poskytnuta všechna relevantní data z provedených metod. Celkové velikosti složek s těmito daty jsou (v pořadí CN-wt, CN-Na41, CN-Cu44) 5,27 GB, 5,21 GB a 4,43 GB, přičemž velikosti FASTQ souborů jsou 5,20 GB, 5,15 GB a 4,37 GB.

Z PDF souboru Analysis_report stojí za zmínění informace o predikci genů. Na základě primární struktury bylo v genomech predikováno (opět v pořadí CN-wt, CN-Na41, CN-Cu44) 7067, 6553 a 7066 genů. Vzhledem k tomu, že v CN H16 genomu z databáze NCBI je 6885 genů [3], část těchto genů je pravděpodobně predikována mylně. Menší množství predikovaných genů u populace CN-Na41 pravděpodobně souvisí s kratším úsekem úspěšně sestavené genomové sekvence CN-Na41. Menší pokrytí u genomové sekvence CN-Na41 je dále diskutováno v kapitole 4.4.2.

Další informace z PDF souboru Analysis_report popisující kvalitu readů a výsledky assembly nebudou zmiňovány, jelikož tyto informace jsou obsažené v kapitolách 4.4.1 a 4.4.2. Informace popisující analýzu mutací zpracovanou společností DNALink nejsou relevantní pro účely této práce, jelikož tyto analýzy mutací byly vždy provedené s ready dané populace oproti její sestavené genomové sekvenci. Popisují tedy pouze “vnitřní variabilitu“ každé populace, nikoliv mutační změny adaptovaných populací oproti původní populaci. Soubory

Analysis_report každé populace jsou dostupné v přílohách 1-3 [DNALink: CN-wt Analysis_Report.pdf].

4.3 Metody použité pro zpracování a analýzu dat

V této kapitole bude probírána většina bioinformatických nástrojů a metod použitých v praktické části této práce. Tyto nástroje a metody byly vybrány a použity pro splnění dvou dílčích úkolů této práce:

- ověření kvality dat, konkrétně readů (FASTQ souborů) a sestavených genomových sekvencí (FASTA souborů)
- sestavení seznamu mutací adaptovaných populací CN-Na41 a CN-Cu44 vůči CN-wt skrze mapování readů adaptovaných populací na sestavenou sekvenci genomu CN-wt

Většina použitých metod na sebe přímo navazuje a jsou tedy uvedené tak, jak chronologicky odpovídají postupu. V navazující kapitole (4.4.) jsou rozebírány výsledky jednotlivých metod.

4.3.1. Analýza kvality readů pomocí FASTQC

Analýza kvality readů je prvním krokem ve zpracování dat z NGS, jelikož kvalita jakýchkoliv dalších analýz přímo záleží na kvalitě readů. Jako nástroj pro tuto analýzu byl zvolen program FASTQC, protože je obecně populární [54], jednoduchý na použití a velice přehledně vizualizuje výsledky. FASTQC byl vytvořen skupinou Bioinformatics Group z Babraham Institute ve Velké Británii za účelem poskytnutí jednoduchého nástroje pro analýzu kvality surových dat z NGS a je zdarma dostupný na jejich stránkách [37].

Vstupní data pro FASTQC musí být ve FASTQ formátu, v našem případě jde tedy dohromady o 6 souborů, jelikož máme R1 a R2 FASTQ soubor pro každou populaci. Analýza každého souboru trvala zhruba 15 minut.

4.3.2. Vyhodnocení kvality assembly pomocí nástroje FASTA statistics z Galaxy

Genomové FASTA sekvence našich tří populací od společnosti DNALink byly sestaveny za použití dat z Illumina short-read sekvenace. Pro stručné vyhodnocení kvality těchto genomových sekvencí byl použit nástroj FASTA statistics [55] z Galaxy. Tento nástroj byl vyvinut v roce 2012 a je aktuálně dostupný pouze z Galaxy [36]. Jde o jednoduchý program, který přečte vložený soubor FASTA a zaznamená základní statistické údaje, jako například průměrnou délku scaffoldů, medián délky scaffoldů, celkový počet N bází a celkový poměr GC.

Vstupní data tohoto programu jsou soubory ve formátu FASTA. V našem případě tedy jde o jeden soubor se zhruba 50 scaffoldy pro každou bakteriální populaci.

4.3.3. Zarovnání scaffoldů vůči referenčnímu genomu pomocí Mauve

Pro porovnání primární sekvence CN-wt s referenčním genomem bude nejdříve nutné vhodně seřadit scaffoldy CN-wt. Assembly totiž správné seřazení scaffoldů nijak nezjistí a pouze je seřadí sestupně dle délky (případně nahodile). Nejvhodnější je seřadit scaffoldy tak, abychom mohli následně porovnat genom s databází NCBI. Pro tento účel se dá použít například program Mauve [56]. Tento program je určený pro porovnávání 2 a více genomových sekvencí najednou, přičemž zarovnávání scaffoldů vůči referenční sekvenci je jedna z jeho vedlejších funkcí. Program Mauve je přehledný, jednoduchý na použití i bez návodu a je zdarma dostupný na stránkách tvůrců – The Darling lab, University of Technology Sydney [57].

Vstupními daty pro Mauve jsou pro naše účely dvě FASTA sekvence. FASTA sekvence se scaffoldy určená na seřazení (CN-wt) a referenční FASTA sekvence podle které budou scaffoldy seřazeny. Jako referenční sekvence byla použita sekvence CN H16 z NCBI [3]. Seřazení bylo hotové za zhruba 10 minut.

4.3.4. Vizualizace FASTA sekvence CN-wt pomocí dot-plotu

Dot-plot je dvourozměrný graf sloužící k porovnání FASTA sekvencí. Na každé ose má vypsanou jednu FASTA sekvenci a pro každou shodu v jejich nukleotidech je v grafu tečka (anglicky „dot“). Poté se z útvarů, které tyto tečky tvoří, dá odvodit podobnost vložených

FASTA sekvencí. Například při vložení dvou identických sekvencí bez repetic se v dot-plotu zobrazí nepřerušená úhlopříčka (začínající v počátku), naopak při vložení dvou FASTA souborů bez jakýchkoliv podobných sekvencí se zobrazí nesouvislé tečky. Pro rychlé a jednoduché vytvoření dot-plotu se dá využít například nástroj D-GENIES [58]. D-GENIES mimo jiné pro zpřehlednění dot-plotu nezobrazuje šum jako osamělé tečky či velice krátké sekvence.

Vstupními daty pro D-GENIES jsou dvě FASTA sekvence libovolných délek a struktur. Vytvoření dot-plotu pro naší sekvenci CN-wt se seřazenými scaffoldy a sekvenci CN H16 z NCBI trvalo několik vteřin.

4.3.5. Mapování readů adaptovaných genomů na FASTA sekvenci CN-wt pomocí BWA

Pro analýzu mutací z našich sad readů je nejdříve nutné přiřadit každou sadu zvlášť k referenční sekvenci pomocí mapování. Pro mapování byl zvolen program BWA, jelikož je obecně účinný při zpracovávání dat z Illumina short-read metod [59] a jeho výstup má vysokou kompatibilitu s programem SAMtools, který byl využit v následujících krocích. Další výhoda BWA je jeho dostupnost a jednoduchost použití skrze portál Galaxy [36]. Tento mapovací program je také dostupný v různých provedeních (specializovaných na ready určité délky) na stránkách autorů [http://bio-bwa.sourceforge.net]. V tomto originálním provedení však není jednoduchý na použití, je nutné jej zkompilovat a ovládá se výhradně skrze příkazovou řádku.

Vstupem pro BWA mapování je referenční sekvence FASTA a sada readů ve formátu FASTQ. V našem případě tedy jde o FASTA sekvenci CN-wt a dva FASTQ soubory s párovými ready pro každou adaptovanou populaci (CN-Na41 a CN-Cu44). Mapování trvalo 2 až 3 hodiny pro každou adaptovanou populaci a výstupní soubory byly formátu BAM.

4.3.6. Filtrování nekvalitních alignmentů pomocí SAMtools

S nižší kvalitou alignemntu se zvyšuje pravděpodobnost, že jemu odpovídající read je přiřazen na špatné místo v referenčním genomu. Při analýze mutací je vhodné tyto (potenciálně) špatné alignmenty odstranit, jelikož můžou hlásit chybné mutace. Pro tento účel je vhodný program SAMtools [39], jelikož umí prohledat soubor BAM a vymazat alignmenty s kvalitou pod

libovolně určenou hranicí. SAMtools je balíček nástrojů určených pro zpracování BAM (případně SAM) souborů, přičemž všechny tyto nástroje jsou dostupné na Galaxy. Konkrétní nástroj použitý pro filtrování alignmentů se nazývá „SAMtools view“.

Vstupem pro SAMtools view je pouze soubor BAM (případně SAM) s danými alignmenty. Jelikož BWA hodnotí kvalitu alignmentů celými čísly mezi 0 a 42, hranice byla nastavena nejprve na 20, při dalším filtrování na 30 a následně na 40. Každé filtrování trvalo zhruba 15 minut.

4.3.7. Seřazení alignmentů a pile-up vůči CN-wt pomocí SAMtools

Pro stanovení mutací v alignmentech je nutné je seřadit podle pořadí bází genomu CN-wt a následně je porovnat s CN-wt vytvořením „pile-up“ souboru. Neshody v tomto porovnání jsou potenciální mutace. Pro seřazení alignmentů v BAM souborech i následné vytvoření pile-up souboru se dá opět použít program SAMtools. Konkrétně se jedná o jeho nástroje „SAMtools sort“ a „SAMtools mpileup“.

Vstupem pro seřazení alignmentů je BAM soubor a vhodná referenční FASTA sekvence. V našem případě byly seřazeny alignmenty filtrovaného BAM souboru CN-Na41 a následně CN-Cu44, vždy vůči sekvenci CN-wt. Následně byl z obou BAM souborů a CN-wt sekvence vytvořen jeden pile-up soubor. Seřazení každého BAM souboru trvalo zhruba 15 minut a vytvoření pile-up souboru zhruba 4 hodiny.

4.3.8. Detekce mutací z pile-up souboru pomocí VarScan 2

K prohledání souboru pile-up se dá použít program VarScan 2 [40] z Galaxy. Tento program projde celý pile-up soubor řádek po řádku, vyhledá řádky s potenciálními mutacemi a vypíše z nich informace relevantní k analýze mutací ve formátu VCF. Analýza pile-up souborů vytvořených pomocí SAMtools je jediná funkce tohoto programu.

Vstupem pro VarScan 2 je pile-up soubor. V našem případě tedy šlo o pile-up soubor dvou adaptovaných populací CN-Na41 a CN-Cu44 s referenční sekvencí CN-wt. VarScan 2 prohledával tento soubor zhruba 8 hodin.

4.4. Výsledky použitých metod

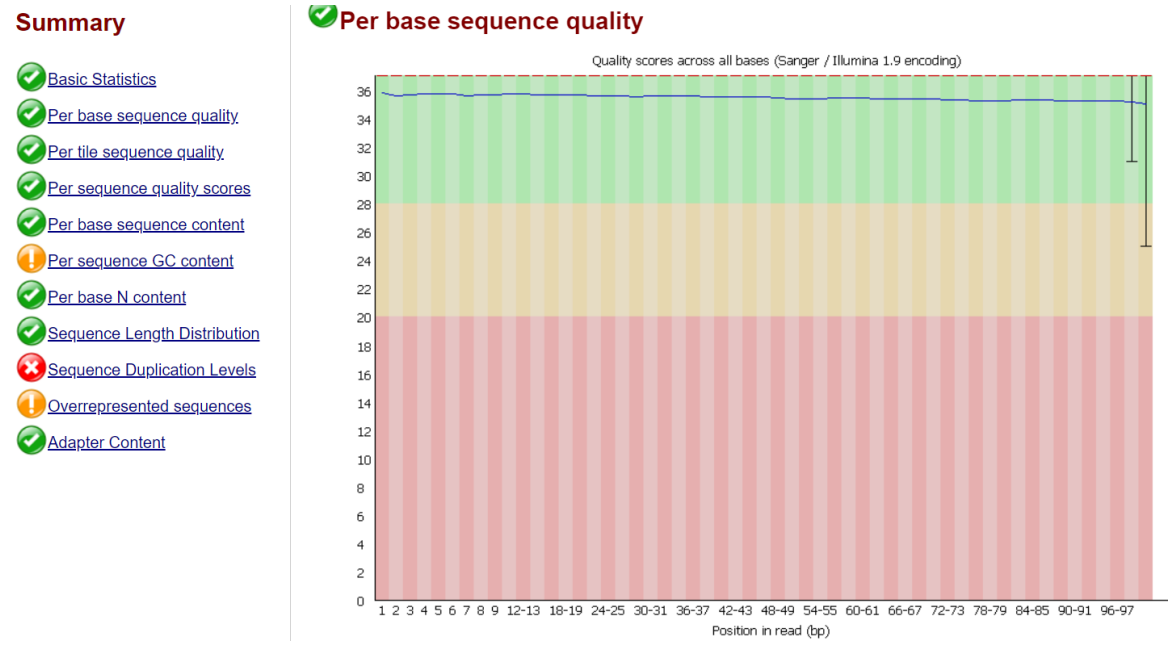
V této kapitole budou rozebírány výsledky metod použitých pro analýzu NGS dat a stanovení mutačních změn. Pořadí, v jakém jsou výsledky rozebírány, odpovídá pořadí metod v kapitole 4.3.

4.4.1. Analýza kvality readů pomocí FASTQC

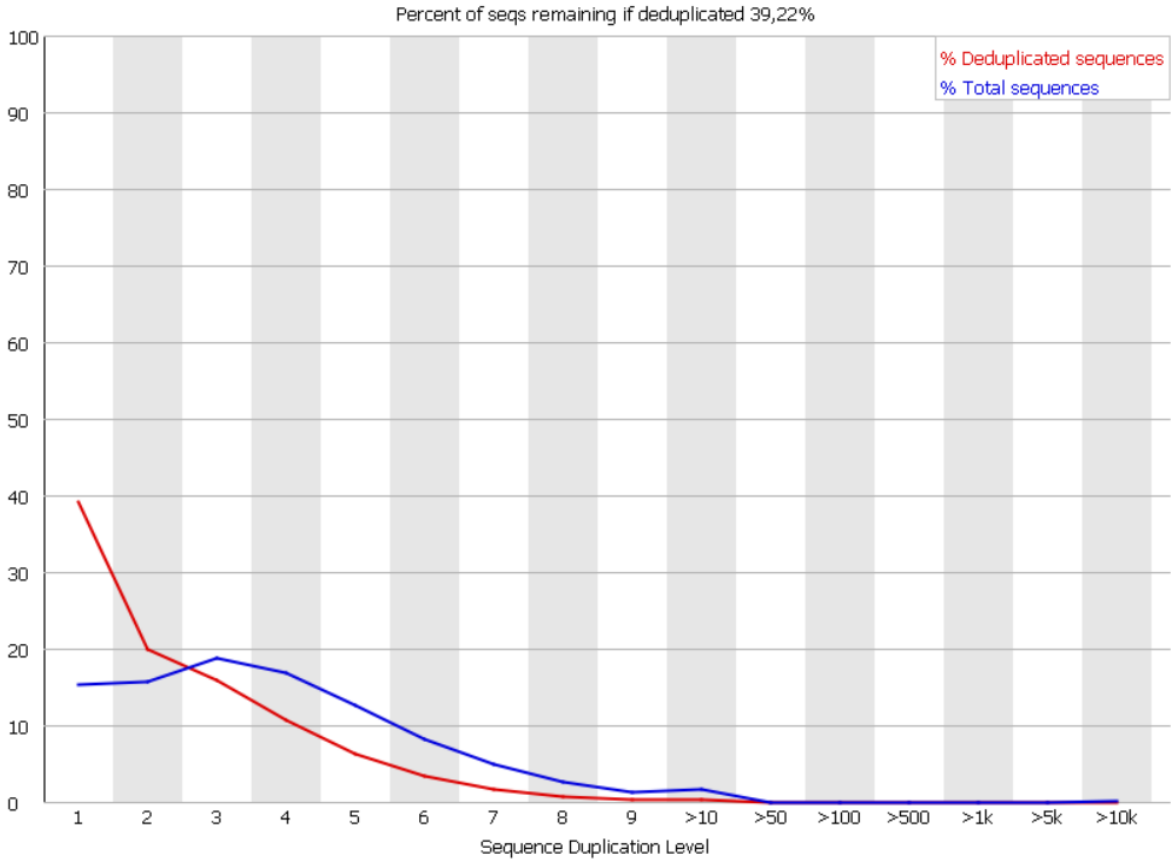
Výsledkem FASTQC analýz byly statistické údaje pro každý FASTQ soubor, rozdělené do 11 kategorií, přičemž každé z těchto kategorií přiřadil program určité orientační ohodnocení (výborné, neutrální nebo záporné). Výsledné soubory se všemi výsledky jsou dostupné v přílohách pod čísly 4-9 [FASTQC: CN-wt R1 Report.html] a dají se otevřít například v internetovém prohlížeči. Výsledky byly pro všech šest souborů podobné a budou shrnuty podle jednotlivých kategorií:

- Basic statistics – kategorie popisující základní informace, například počet readů, jejich délka a poměr GC. Pro všechny tři populace vyšly dle očekávání délky readů 101 a poměry GC zhruba 66 procent. Počet readů vyšel pro populaci CN-Cu44 přibližně 85 milionů, pro ostatní dvě populace přibližně 99 milionů.
- Per base sequence quality [viz Obrázek 6] – v této kategorii je graf popisující průměrnou kvalitu báze v závislosti na pozici v readu. Pro všechny naše soubory se průměrná kvalita na všech pozicích držela nad 35, což udává více než 99,9 % pravděpodobnost správného určení báze. V R2 readech všech tří populací se ke konci readu kvalita nepatrně snižovala.
- Per tile sequence quality – tato kategorie ukazuje průměrnou kvalitu báze v závislosti na pozici klastru v průtokové cele. Pro naše data byla kvalita všude naprosto stejná.
- Per sequence quality scores – graf znázorňující distribuci skóre kvality v bázích. Pro báze v našich souborech byla vždy zdaleka nejčastější kvalita 36, poté kvalita 35 a 37.
- Per base sequence content – graf znázorňující poměr bází v readech v závislosti na pozici v readu. Pro všechna naše data byl poměr dle očekávání konstantní.
- Per sequence GC content – tato kategorie znázorňuje průměrný poměr GC v každém readu. U R1 souborů vyšel průměrný poměr GC dle očekávání, u R2 souborů však ne (z důvodů diskutovaných níže).

- Per base N content – stejný graf jako v kategorii “Per base sequence content“, znázorňující však pouze N báze. Pro všechny soubory vyšel poměr N bází zdánlivě nulový.
- Sequence length distribution – graf znázorňující distribuci délek readů. Pro všechny soubory byl graf stejný s jediným peakem na délce 101.
- Sequence duplication levels [viz Obrázek 7] – tato kategorie znázorňuje počet readů, které jsou z úplně stejného místa. Pro všechny naše soubory vyšel tento graf nepravidelný a s červeným varováním od FASTQC.
- Overrepresented sequences – v této kategorii program vypisuje sekvence, které se často opakují a tvoří více než 0,1 % informace ve FASTQ souboru. Ve všech R2 souborech byla nalezená artefaktní sekvence 50 guaninů tvořící zhruba 0,3 % celého souboru.
- Adapter content – v této kategorii jsou případně vypsány sekvence které FASTQC rozezná jako chybně nasekvenované adaptéry z NGS. Pro všechny naše soubory byla tato kategorie prázdná.



Obrázek 6 - Graf z kategorie Per base sequence quality pro R2 ready populace CN-Cu44. Modrá čára znázorňuje aritmetický průměr kvality v závislosti na dané pozici. Černé čáry na pravé straně znázorňují zvýšený rozptyl kvality na těchto pozicích. Na levé straně jsou vidět jednotlivé kategorie a orientační ohodnocení přiřazené programem FASTQC.



Obrázek 7 - graf z kategorie Sequence duplication levels pro R2 ready populace CN-Cu44. FASTQC udělil v této kategorii červené varování kvůli časté totožnosti readů.

Z těchto výsledků se dá vyvodit několik klíčových informací. Například že průměrný počet readů připadajících na jednu bázi z genomu CN H16 je pro každou naší populaci více než 1000, přičemž pro analýzu bodových mutací obecně stačí 15 až 20 a pro analýzu inzerčních/delečních mutací stačí 300 [60]. Dále je evidentní, že průměrná pravděpodobnost správného určení báze je velice vysoká, poměr N bází k ACGT bázím se blíží k nule, průměrné rozložení bází v readech je rovnoměrné a nejsou přítomny žádné adaptérové sekvence z NGS, což jsou vše velice pozitivní informace. Ve všech souborech je ale vysoká míra duplikátních readů a v R2 souborech jsou (50 bází dlouhé) opakované guaninové sekvence, které jsou chybně zaznamenané pravděpodobně z důvodu vysokého obsahu GGC (CCG) tripletů, případně inverzních repetic v určitých místech sekvenovaného genomu [61]. Kvůli těmto sekvencím vyšel v R2 souborech průměrný poměr GC na sekvenci nad očekávánou hodnotu.

V dalším postupu analýzy mutací však guaninové sekvence nebudou komplikací, jelikož při mapování je algoritmus nepoužije. Úroveň duplikace by taktéž neměla být komplikací, jelikož i přes očekávání duplikáty obecně nezkreslují výsledky při hledání mutací [60]. Vysoká úroveň duplikace mimo jiné pravděpodobně souvisí s velkým poměrem readů připadajících na jednu bázi. Celkově se tedy kvalita readů jeví jako vhodná pro analýzu mutací.

4.4.2. Vyhodnocení kvality assembly pomocí nástroje FASTA statistics z Galaxy

Výsledkem nástroje FASTA statistics byly stručné textové soubory se základními statistikami popisujícími vložené FASTA sekvence. Výsledky byly zaznamenány do tabulky, viz Tabulka 2.

FASTA sekvence	Poměr GC	Počet N bází	Délka sekvence (včetně N bází)	Počet scaffoldů	Medián délky scaffoldu
CN-wt	66,4 %	5436	7 369 813 bp	50	68 403 bp
CN-Cu44	66,4 %	5241	7 369 628 bp	46	71 740 bp
CN-Na41	66,6 %	2301	6 926 421 bp	41	70 832 bp

Tabulka 2 - výsledky tří analýz genomových sekvencí pomocí nástroje Fasta statistics. Délka sekvence je součet délek scaffoldů.

Tyto statistiky dávají dobrou orientační představu o věrohodnosti jednotlivých sestavených genomů. Genom CN H16 má délku 7 414 561 bp a poměr GC 66,36 % [3], čemuž se první dva genomy velice blíží. CN-Na41 je ale o zhruba 440 000 bp kratší, což je vzhledem ke struktuře evolučního experimentu téměř nemožné, tedy jde o chybu NGS. Jelikož párové ready jsou dle naší předešlé analýzy velice kvalitní, chyba se musela stát při assembly. V CN-Na41 je také o zhruba 3000 N bází méně, je tedy možné že "chybějící část genomu" měla příliš velkou koncentraci N bází na to, aby byla algoritmem sestavena. Pro analýzu mutací však bude z těchto tří genomových sekvencí použita pouze sekvence CN-wt, tím pádem jsou chyby v CN-Na41 nevýznamné. Dle těchto výsledků je genomová sekvence CN-wt vhodná pro analýzu mutací. Přítomnost N bází v této sekvenci bude v dalších analýzách zohledněna.

4.4.3. Seřazení scaffoldů vůči referenčnímu genomu pomocí Mauve

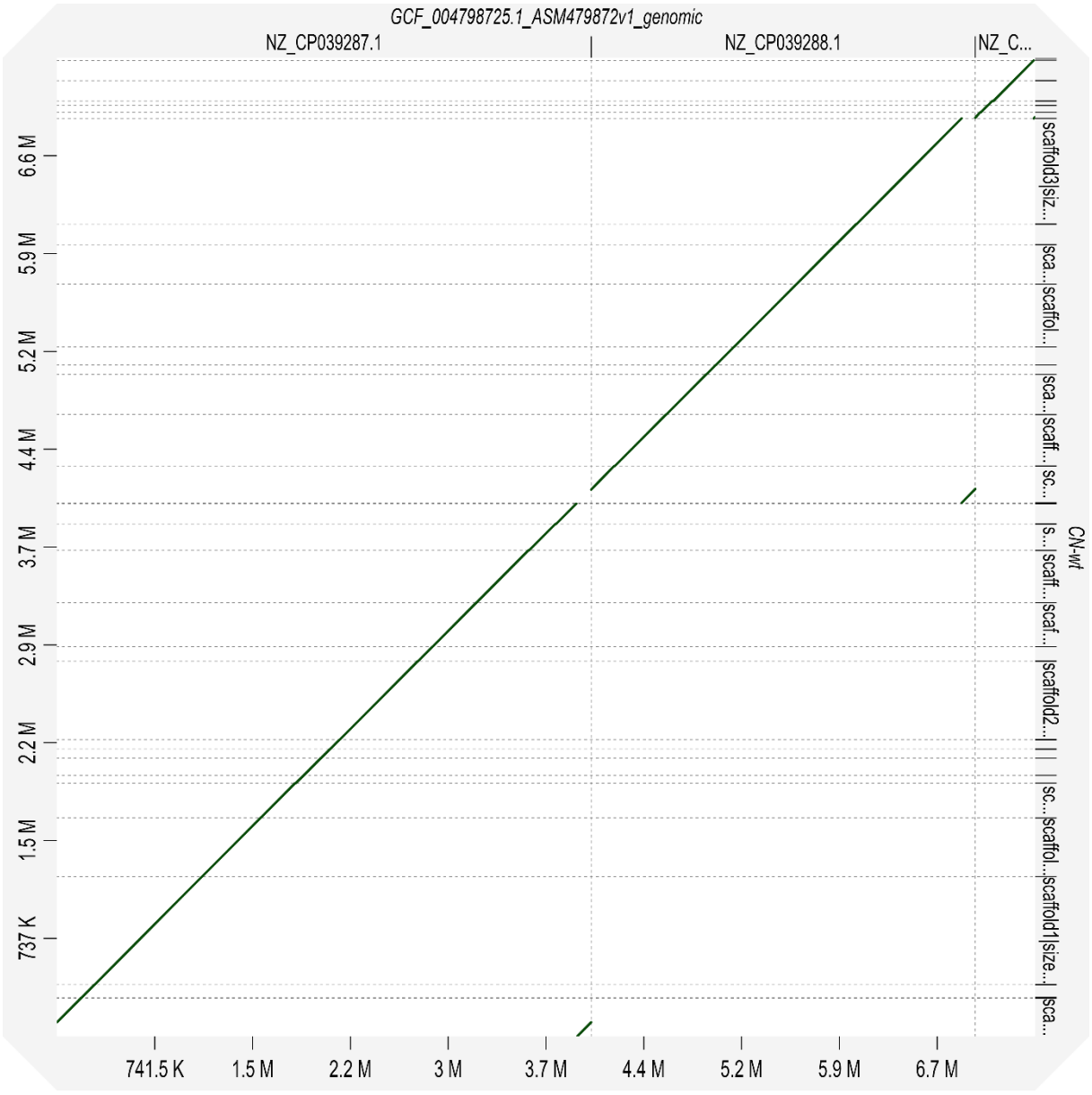
Výsledkem seřazení scaffoldů CN-wt byl nový FASTA soubor se seřazenými scaffoldy. Pro vyhodnocení úspěšnosti seřazení a porovnání seřazené sekvence s referenční je třeba využít nějakou další metodu.

4.4.4. Vizualizace FASTA sekvence CN-wt pomocí dot-plotu

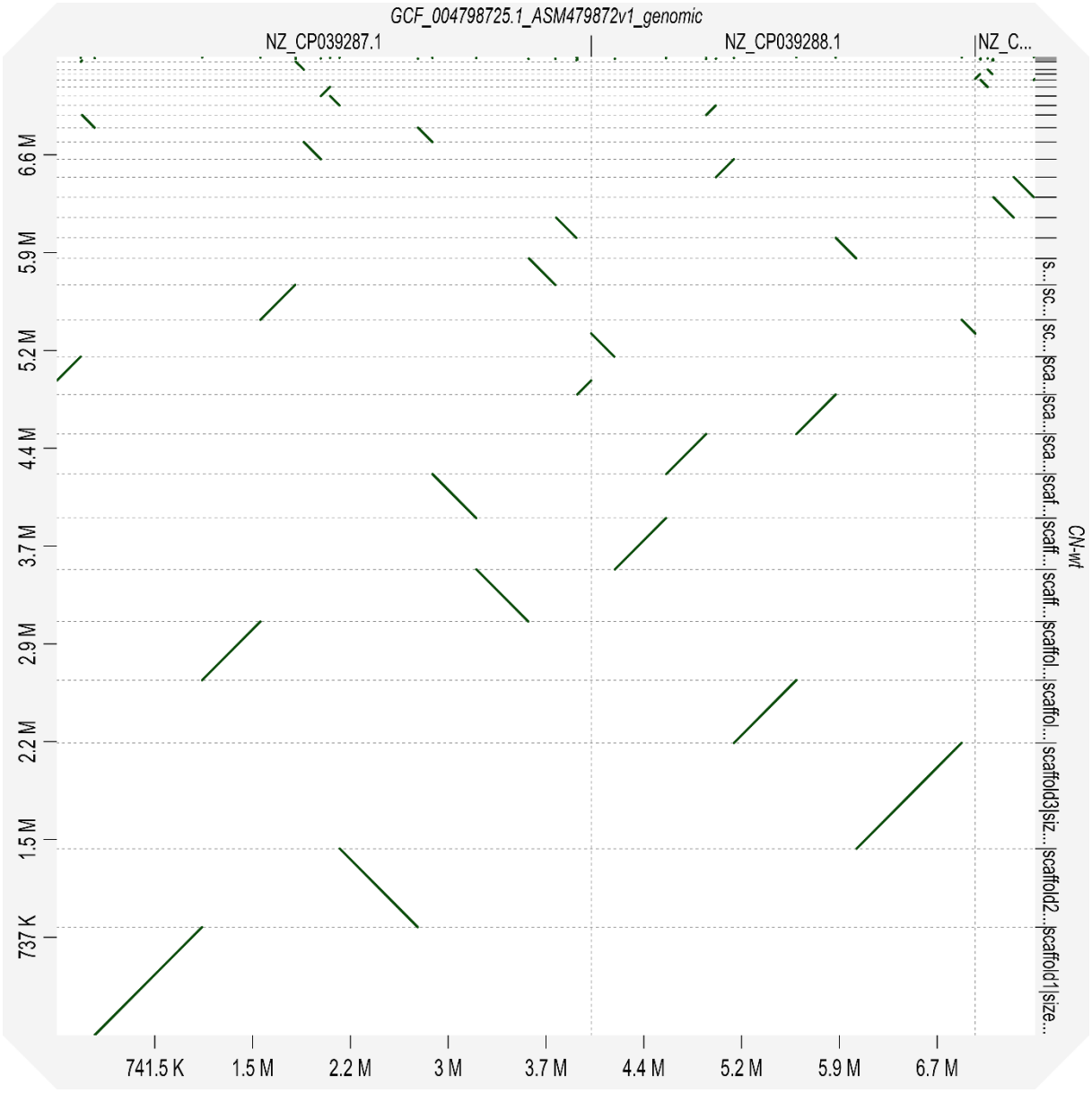
Výsledkem nástroje D-GENIES byl obrázek dot-plotu tvořený převážně diagonálními úsečkami, viz Obrázek 8. Z obrázku je na první pohled vidět, že vložené sekvence jsou značně podobné. Malé osamělé úseky uprostřed a dole nejsou na očekávaných pozicích kvůli struktuře scaffoldů CN-wt – na vertikální ose je vidět, že tyto osamělé úseky jsou součástí scaffoldů obsahujících více sekvencí z různých míst na genomu z NCBI, což je problém neřešitelný přeřazováním scaffoldů. Chyba tohoto typu je pro Unicycler (program použitý pro assembly firmou DNALink) při zpracovávání krátkých readů častá [62]. Pro vyhodnocení úspěšnosti seřazení může posloužit dot-plot se sekvencí CN-wt před seřazením scaffoldů, viz Obrázek 9.

Ačkoliv se sekvence (seřazená sekvence CN-wt a sekvence z NCBI) při pohledu na dot-plot zdají totožné v primární struktuře, malé případné rozdíly v rozsahu desítek až stovek nukleotidů, jsou v tomto měřítku bohužel neviditelné. V dalším postupu bude tento fakt

zohledněn a určité úseky sekvence CN-wt budou odlišnou metodou detailně porovnány se sekvencí z NCBI pro případnou detekci chyb assembly.



Obrázek 8 - Dot-plot se sekvencí z NCBI na horizontální ose a seřazenou CN-wt sekvencí na vertikální ose. Jednotky obou os jsou [bp]. Horizontální tečkované čáry ukazují konce scaffoldů CN-wt a vertikální tečkované čáry ukazují konce chromozomů sekvence CN H16 z NCBI.



Obrázek 9 - Dot-plot se sekvencí z NCBI na horizontální ose a neseřazenou CN-wt sekvencí na vertikální ose. Jednotky os jsou [bp].

4.4.5. Mapování readů adaptovaných genomů na FASTA sekvenci CN-wt pomocí BWA

Výsledkem mapování pomocí BWA byly BAM soubory adaptovaných populací. Oba soubory měly v této fázi velikost zhruba 3 GB a obsahovaly určité množství nekvalitních alignmentů.

4.4.6. Filtrování nekvalitních alignmentů pomocí SAMtools

Výsledkem filtrování byl BAM soubor s alignmenty kvality 40 a více pro obě adaptované populace. Pomocí SAMtools byly stručně vyhodnoceny BAM soubory před a po filtrování, vybrané výsledky byly poté zaznamenány do tabulky, viz Tabulka 3.

BAM soubor	CN-Na41 před filtr.	CN-Na41 po filtr.	CN-Cu44 před filtr.	CN-Cu44 po filtr.
Počet readů	99 312 334	94 533 774	83 696 094	80 069 024
Počet namapovaných readů	97 656 155	94 525 545	82 744 490	80 062 390
Správně spárované ready	96 450 288	94 525 545	81 783 520	80 062 390
Osamělé ready	976 982	0	733 776	0
Páry na různých scaffoldech	241 816	13 493	239 720	10 961

Tabulka 3 - Statistika filtrování namapovaných readů

Z prvních dvou řádků tabulky je evidentní že BWA algoritmus využil naprostou většinu dostupných readů. Toto je očekávaný výsledek, jelikož ready jsou velice kvalitní a pochází z genomů podobných CN-wt. Správně spárovanými ready jsou myšleny ty páry readů, které po mapování mezi sebou mají korektní vzdálenost. Tato vzdálenost je pro všechny páry ideálně konstantní a je určená podmínkami NGS. Ačkoliv tato vzdálenost není nikde v souborech explicitně uvedena, BWA jí odhadne na konci mapování na základě svých výsledků a vyhodnotí, zda namapované páry tuto vzdálenost mezi sebou mají a jsou tedy namapovány korektně. Jelikož naprostá většina našich namapovaných readů byla správně spárována, můžeme mapování považovat za úspěšné. Osamělé ready jsou ready, které nebyly spárovány. Jsou tedy pravděpodobně chybně namapované a je žádoucí, aby jejich počet byl co nejnižší. Páry na různých scaffoldech jsou páry, z nichž jeden read je namapován na jednom scaffoldu CN-wt a druhý na jiném. Jelikož naše CN-wt scaffoldy jsou seřazené, nemusí nutně jít o chybu. Jelikož filtrované BAM soubory nyní obsahují pouze alignmenty velice vysoké kvality (40 až 42) a většina artefaktů byla odstraněna, BAM soubory se dají považovat za kvalitní, a tedy vhodné pro analýzu mutací.

4.4.7. Seřazení alignmentů a pile-up vůči CN-wt pomocí SAMtools

Výsledkem těchto dvou kroků byl pile-up soubor s informacemi o každé pozici referenční sekvence a o readech relevantních k této pozici. Pro analýzu mutací jsou důležité hlavně počty readů na každé pozici, které se shodují s referenční bází nebo alternativní báze navrhované ready, které se s referencí neshodují. Ačkoliv tento pile-up soubor obsahuje všechny hledané informace o mutacích, kvůli jeho velikosti (zhruba 32 GB) bylo ruční hledání mutací takřka nemožné.

4.4.8. Detekce mutací z pile-up souboru pomocí VarScan 2

Výsledkem této analýzy byl VCF soubor s řádky popisujícími jednotlivé potenciální mutace, jež měly alespoň 5 % frekvenci v jedné z populací. Soubor měl 554 řádků popisujících bodové mutace a 23 řádků popisujících deleční/inzerční mutace. Na první pohled však bylo jasné, že většina řádků popisuje mutace na N bázích genomu CN-wt, které bylo nutné před dalším rozborem mutací odstranit. Tento VCF soubor je dostupný v příloze 10 [VarScan 2 Report.vcf] a dá se otevřít v libovolném textovém editoru.

4.5. Finalizace výsledků analýzy mutací

Výsledkem analýzy mutací mapováním pomocí programu BWA a následným hledáním mutací pomocí SAMtools byl seznam 577 potenciálních mutací [VarScan 2 Report.vcf]. Před analýzou těchto mutací bylo však nutné odstranit chybné či irrelevantní mutace. Mezi tyto patří zejména mutace na N bázích a mutace způsobené chybou v mapování programu BWA.

Jelikož náš referenční genom CN-wt obsahuje 5436 N bází, byl předpoklad že na těchto pozicích budou mylně hlášeny mutace. V první verzi VCF souboru bylo nalezeno 438 mutací na N bázích (426 bodových, 12 inzerčních/delečních). Těchto mutací nebylo nahlášeno větší množství pravděpodobně proto, že většina N bází v CN-wt je v repetitivních sekvencích (typu NNNNN…), na které algoritmus BWA nemohl korektně žádné ready namapovat. Navíc všechny pozice s N bázemi měly velice nízké pokrytí ready a téměř 100 % frekvenci alternativní báze. Z těchto důvodů byly všechny neshody tohoto typu byly odstraněny.

Pozice zbylých mutací (128 bodových, 11 inzerčních/delečních) byly pomocí nástroje NCBI-Blast [63], dostupného na internetových stránkách NCBI [64]. asociovány s odpovídajícími pozicemi na NCBI CN H16 genomu [3]. Většina těchto potenciálních mutací (130/139) se nacházela v genových oblastech. O většině těchto genových oblastí se na NCBI vyskytují určité informace, které byly zaznamenány. Seznam těchto mutací ve formě excelové tabulky je dostupný v příloze 11 [Seznam 139 mutací.xlsx].

Po odstranění mutací na N bázích a asociování zbylých 139 potenciálních mutací s dostupnými informacemi z NCBI byly odstraněny další chyby a mutace. Pro charakterizaci odchylek od průměrných hodnot byla v této kapitole použita směrodatná odchylka. Následující typy chyb, případně mutací, byly odstraněny v tomto pořadí:

- Chybně hlášené bodové mutace vinou nekorektních alignmentů – celkem 33 ze 139
    - BWA, algoritmus použitý pro mapování, považuje korektní vzdálenost mezi párovými ready jako prioritní informaci při tvorbě alignmentů. Vinou tohoto postupu může nastat chyba, když jeden read z páru se namapuje na určité místo perfektně a jeho partner se namapuje v korektní vzdálenosti od něj s dostatečně málým množstvím neshod. Následně při hledání mutací jsou tyto neshody považovány programem za bodové mutace [65]. Výsledkem těchto chyb tedy jsou skupiny těsně sousedících bodových mutací, vždy s přibližně 50 % frekvencí [66]. Tato frekvence je určená poměrem chybných a korektních alignmentů pro dané místo, který za daných podmínek bývá stejný.
    - V našich datech se tyto chyby vyskytovaly v uskupení průměrně 4,7 ± 1,5 bodových mutací na scaffoldech 40, 41, 42, 43, 44, 45 a 47. Průměrná délka těchto scaffoldů je (845,7 ± 151,9) bp. Nízká délka scaffoldu obecně implikuje chyby, například někteří bioinformatici při analýze mutací po assembly odstraňují scaffoldy kratší než 1000 bp [67]. Chybné alignmenty tedy mimo jiné pravděpodobně vznikly vinou chyb v kratších scaffoldech.
- Artefaktní bodové mutace specifické pro CN-Cu44 – 49 ze zbývajících 106
    - Tyto mutace v CN-Cu44 readech byly označeny za artefakty primárně z důvodu podezřele nízkého pokrytí – průměrně 26 ± 7 readů. Jelikož ostatní mutace obou adaptovaných populací měly průměrné pokrytí 1406,6 ± 699,3 readů, takto nízké pokrytí implikuje značnou nejistotu. Dále tyto mutace měly poměrně nízkou frekvenci, v průměru (9,1 ± 3,2) %, přičemž 5 % je potřebných pro

zaregistrování algoritmem hledajícím mutace. Ready CN-Na41 na těchto pozicích měly průměrné pokrytí 1194 ± 338 na jednu bázi a nulovou frekvenci u jakékoliv z těchto mutací. Vzhledem k těmto faktům jde pravděpodobně o chyby způsobené nevhodným alignmentem určitých CN-Cu44 readů.

- Mutace přítomné ve všech třech populacích ve stejné frekvenci – 42 z nynějších 57
    - Mutace tohoto typu nejsou výsledkem evoluční adaptace na prostředí přítomná v experimentu, a proto nejsou relevantní. Stejnou frekvencí je myšlena frekvence v populaci CN-wt ± 1 %.
    - Pro jejich stanovení bylo nutné zopakovat postup analýzy mutací s ready CN-wt.
- Inzerční/deleční mutace s nízkým pokrytím – 4 ze zbývajících 15
    - Všechny 4 inzerční/deleční mutace přítomné ve zbývajících 15 mutacích měly velice nízké pokrytí, v průměru 49 ± 27 readů na mutaci. Jelikož průměrné pokrytí v rámci celého mapování je pro každou populaci větší než 1000 readů na bázi, takto nízké pokrytí opět naznačuje možnost chybného hlášení mutací způsobeného nekorektním mapováním. Mutace však byly odstraněny primárně, protože pokrytí potřebné pro korektní stanovení inzerčních/delečních mutací je dle dostupných zdrojů zhruba 300 readů na bázi [60].

Výsledné mutace byly společně s relevantními informacemi zpracovány do přehledné tabulky, viz Tabulka 4. Pro vyjádření frekvencí mutací v tabulce byla použita průměrná frekvence z kódujícího a pracovního vlákna, jelikož tyto frekvence se ve zmíněných mutacích lišily minimálně. Zmíněné funkce většiny těchto genů jsou kombinacemi informací z bioinformatických databází odvozených (programy) na základě funkcí homologních genů, a tedy pouze orientační.

#	scaffold	pozice	ref. báze	mutace	f(Na41)	f(Cu44)	f(wt)	jednotka kódovaná daným úsekem
1	scaffold1	342094	A	C	13,66%	10,58%	7,17%	transmembránový protein rodiny AEC
2	scaffold20	18330	G	A	10,20%	0,14%	3,65%	protein obsahující doménu DUF2063
3	scaffold6	221432	C	T	7,59%	0,08%	0,00%	podjednotka glykolát oxidázy GlcE
4	scaffold6	221706	C	T	0,00%	9,24%	0,00%	
5	scaffold6	222008	A	C	0,11%	69,37%	0,00%	
6	scaffold7	54988	C	G	0,08%	64,85%	0,00%	hypotetický protein
7	scaffold3	464225	G	A	87,03%	0,00%	0,00%	transkripci regulující protein rodiny LysR
8	scaffold3	464714	G	A	10,71%	0,00%	0,00%	
9	scaffold35	338	C	T	17,80%	23,89%	18,88%	23S podjednotka ribosomální RNA
10	scaffold35	1782	T	A	5,26%	7,22%	6,45%	nekódující oblast za 23S rRNA genem
11	scaffold46	435	C	T	49,43%	1,69%	42,93%	transketoláza

Tabulka 4 - Tabulka znázorňující stanovené mutační změny v populacích po adaptaci na daná prostředí. Sloupce "Ref. báze" a "Mutace" popisují varianty bází kódujícího vlákna na dané pozici. Sloupce f(X) popisují frekvenci mutace v readech dané populace. Na žádné pozici nebyla stanovena více než jedna mutace. Uvedené kódované jednotky jsou ve většině případů pouze orientační informací stanovenou programy na základě homologie genů.

4.6. Rozbor stanovených mutačních změn

V této kapitole budou jednotlivě či po skupinách dále rozebrány mutace z výše umístěné tabulky [Tabulka 4] dle jejich pořadí. Rozbor byl proveden především na základě informací o relevantních úsecích genomu CN H16 dostupných z vědeckých článků a jiných ověřených zdrojů. V případech některých mutací byly dostupné zdroje značně omezené. Zmíněné lokace mutací jsou vztažené k první bázi ve scaffoldu, případně k první bázi v genu ve směru transkripce, kde tyto báze mají lokaci 1. Konkrétní zmíněné báze vždy popisují kódující vlákno.

4.6.1. #1 Bodová mutace A→C na pozici 342094 ve scaffoldu 1

Gen, v jakém se tato mutace nachází, kóduje transmembránový protein rodiny AEC. Mutace se nachází na bázi 424, přičemž celková délka genu je 2904 bází.

Rodina AEC (Auxin Efflux Carrier) proteinů je asociovaná především s rostlinami, ve kterých slouží k transmembránovému přenosu auxinu (rostlinného hormonu). U bakterií však existují homologní proteiny zprostředkovávající přenos aminokyselin přes buněčnou membránu [68], přičemž akumulace aminokyselin (a dalších látek) je mimo jiné primární reakcí na osmotický stres [69].

Bodová mutace stanovená ve všech našich CN populacích mění ACC triplet (Thr) na CCC triplet (Pro). Tato mutace se vyskytovala v určité úrovni v původní populaci (7,17%), se zvýšenou frekvencí u populace Cu44 (10,58%) a s nejvyšší frekvencí u populace Na41 (13,66%). Možné vysvětlení je, že mutace souvisí s adaptací na osmotický stres a má určitý pozitivní vliv na fungování kódovaného proteinu v prostředí s osmotickým stresem.

4.6.2. #2 Bodová mutace C→T na pozici 18330 ve scaffoldu 20

Popis tohoto genu na NCBI je omezený na informaci, že kóduje protein obsahující doménu DUF2063. Celková délka genu je 861 nukleotidů, přičemž mutace je lokalizována na nukleotidu 722. Délka domény DUF2063 je 82 aminokyselin [70], což odpovídá 246 bázím. Není tedy jasné, zda se naše mutace nachází přímo v úseku kódujícím tuto doménu.

Proteiny obsahující doménu DUF2063 mají strukturu podobnou transkripčním faktorům, což je jejich předpokládaná funkce [70]. Mutace v našich populacích mění TGT triplet (Cys) na TAT triplet (Tyr) a vzhledem ke zvýšené frekvenci u populace Na41 (a předpokládané funkci genu) pravděpodobně ovlivňuje regulaci transkripce v kontextu zvýšeného osmotického stresu [71].

4.6.3. #3; #4; #5 Bodové mutace na pozicích 221432 až 222008 ve scaffoldu 6

Gen glcE, kódující podjednotku glykolát oxidázy, má celkovou délku 1104 bází a po adaptaci na stresové podmínky v něm byly stanoveny tři nové bodové mutace na pozicích 367, 656 a 943. Tyto mutace mění dle pořadí triplety ATG (Met), ACC (Thr) a CAG (Glu) na CTG (Leu), ACT (Thr) a TAG (stop kodon).

Glykolát oxidáza katalyzuje oxidaci glykolátu na glyoxylát, který v glyoxylátovém cyklu za katalýzy malátsyntázou reaguje s Acetyl-CoA za tvorby malátu a CoA. Glyoxylátový cyklus v bakteriích slouží primárně ke tvorbě karbohydrátů v prostředích bez dostupné fruktózy/glukózy. Jelikož naše bakteriální populace rostly v prostředí s fruktózou, neměly důvod tuto metabolickou dráhu nijak významně využívat. Tato dráha také využívá Acetyl-CoA, jehož koncentrace má přímý vliv na akumulaci PHA (konkrétně poly-3-hydroxybutyrátu) [43].

Jelikož akumulace PHA pomáhá CN H16 vypořádat se se stresovými podmínkami, je pravděpodobné, že mutace stanovené v genu glcE způsobily inhibici glykolát oxidázy a tím zvýšily tok Acetyl-CoA do metabolické dráhy syntézy PHA. Toto se ovšem netýká mutace #4 ACC→ACT, jelikož obě varianty jejích výsledných kodonů kódují Threonin. Rozšíření této neutrální mutace na 9,24 % populace Cu44 bylo pravděpodobně způsobeno sepětím této mutace s jinou, pozitivní mutací.

4.6.4 #6 Bodová mutace C→G na pozici 54988 ve scaffoldu 7¨

O genovém úseku obsahujícím tuto mutaci nejsou dostupné žádné informace kromě jeho délky (1179 bp). Mutace GCG→GAG (Ala→Glu) na pozici 929 přítomná v 64,85 % variantách tohoto genu v populaci Cu44 naznačuje určitou selekci této mutace v kontextu adaptace na Cu^{2+} ionty. Nejsou však žádné biochemické (či jiné) důkazy pro existenci hypotetického proteinu kodovaného touto sekvencí.

4.6.5. #7; #8 Bodové mutace na pozicích 464225 a 464714 ve scaffoldu 3

Gen s těmito dvěma mutacemi kóduje transkripční regulátorový protein rodiny LysR. Mutace CGC→CAC (Arg→His) je v genu na pozici 752 a mutace CGG→CAG (Arg→Glu) na pozici 1241, přičemž celý gen má délku 1380 bp.

Proteiny rodiny LysR patří mezi nejčastější regulátory transkripce u prokaryot a mají uplatnění mimo jiné při regulaci metabolismu. Vzhledem ke konzervované struktuře těchto genů [72] měly mutace #7 a #8 vyvinuté v populaci Na41 pravděpodobně pozitivní efekt na její biologickou zdatnost skrze vliv na regulaci transkripce při osmotickém stresu [71].

4.6.6. #9; #10 Bodové mutace na pozicích 338 a 1782 ve scaffoldu 35

Mutace #9 a #10 se nachází v operonu strukturních RNA genů kódujících (ve směru transkripce) rRNA 16S podjednotku, tRNA-Ile, tRNA-Ala, rRNA 23S podjednotku a rRNA 5S podjednotku. Mutace #9 (C→T) se nachází v genu rRNA 23S podjednotky na pozici 1786 z celkové délky genu 2904 bp a mutace #10 (T→A) se nachází 56 bází za tímto genem.

Úsek mezi 23S a 5S rRNA genem je u bakterií poměrně konzervovaný [73] a použitelný na identifikaci odlišných rodů bakterií [74]. Vzhledem však k nízké změně ve frekvenci mutace #10 je souvislost této mutace s adaptací na prostředí experimentu nepravděpodobná.

Gen kódující podjednotku 23S rRNA je značně konzervovaný a bodové mutace v něm mohou způsobit například rezistenci proti antibiotikům blokujícím translaci proteinů [75][76]. Ačkoliv rezistence proti antibiotikům se našeho experimentu netýká, navýšení frekvence mutace #9 v populaci Cu44 o zhruba 5 % naznačuje určitou evoluční adaptaci. Zvýšení koncentrace těžkých kovů, jako například mědi, může způsobit chyby ve skládání proteinů a jejich agregaci [77]. Jelikož 23S rRNA podjednotka je důležitá při skládání proteinů po translaci [78], je možné že mutace #9 pozitivně ovlivňuje tuto aktivitu 23S podjednotky v prostředí s vyšší koncentrací Cu^{2+} iontů, a proto došlo ke zvýšení její frekvence v populaci Cu44.

4.6.7. #11 Bodová mutace C→T na pozici 435 ve scaffoldu 46

Mutace #11 se vyskytuje na bázi 795 genu kódujícího enzym transketolázu celkové délky 2013 bp. Jde o mutaci, která mění CAG triplet (Glu) na TAG triplet (stop kodon). Vzhledem k lokaci a charakteru této mutace je její nejpravděpodobnější následek předčasné ukončení transkripce a zabránění tvorbě enzymu transketolázy.

Transketolázu bakterie využívají v rámci pentóza-fosfátového cyklu pro katalýzu tvorby glyceraldehyd-3-fosfátu z ribosa-5-fosfátu či xylulosa-5-fosfátu. Glyceraldehyd-3-fosfát je v glykolýze následně převeden na pyruvát, který je dekarboxylován na Acetyl-CoA. Skrze její vliv na koncentraci Acetyl-CoA má tedy transketoláza vliv na syntézu PHA. Například transformace producentů PHA amplifikovaným genem kódujícím transketolázu vede ke zvýšení akumulace PHA a růstu bakterie [79].

Pokles ve frekvenci TAG mutace z CN-wt populace na CN-Cu44 by se tedy dal vysvětlit selekčním tlakem pro mutanty s lepšími schopnostmi akumulace PHA (zprostředkovaným Cu^{2+} ionty). Možné vysvětlení však také je chybné hlášení mutací kvůli nekorektnímu mapování readů na scaffold 46 v případě populací CN-Na41 a CN-wt, ze stejných důvodů jako v případě chybných mutací ve scaffoldech 40, 41, 42, 43, 44, 45 a 47. Jelikož na pozici této mutace připadá v populacích CN-Na41 a CN-wt přibližně dvakrát více readů (700 a 934) než v populaci CN-Cu44 (473), je možné že přibližně polovina těchto readů pochází z jiné oblasti genomu a jde o chybu mapování.

4.7. Diskuse důsledků mutačních změn na produkci PHA

V populacích CN-Cu44 a CN-Na41 nebyly nalezeny žádné mutační změny v genech kódujících enzymy metabolické dráhy syntézy PHA v phaCAB operonu. Dále nebyly nalezeny žádné mutace v genech phaP kódujících proteiny stabilizující granule PHA. Pro ověření nepřítomnosti těchto mutací v původním genomu byly porovnány části FASTA sekvence CN-wt se sekvencemi těchto genů z NCBI, přičemž ve všech případech se potvrdila úplná shoda našich sekvencí z CN-wt s databázovými sekvencemi. Výsledky těchto porovnání pomocí NCBI-Blast jsou dostupné v přílohách 12 a 13 [NCBI-Blast report - phaCAB + phaR porovnáno s CN-wt.txt, NCBI-Blast report - phaP vs CN-wt.txt].

Ačkoliv tedy nebyly stanoveny žádné mutace v genech s jednoznačným důsledkem na tvorbu PHA, byly stanoveny mutace v genech na různých úrovních souvislosti s produkcí PHA. Například mutace #2 u CN-Na41 v genu kódujícím protein obsahující doménu DUF2063, který dle předpokladů souvisí s regulací transkripce [65], by mohla určitým způsobem ovlivňovat citrátový cyklus, jelikož exprese jeho enzymů je v reakci na osmotický tlak regulována hlavně transkripcí [71]. Mutace #7 a #8 u populace CN-Na41 jsou taktéž v genech kódujících proteiny sloužící k regulaci transkripce a mohly by tedy mít podobný efekt. Tyto mutace jsou relevantní k produkci PHA, protože citrátový cyklus, konkrétně koncentrace acetyl-CoA a koncentrace NADPH přímo souvisí s akumulací PHB u CN H16 [43].

Mutace #3 u populace CN-Na41 a mutace #5 u populace CN-Cu44 v genu kódujícím podjednotku glykolát oxidázy by taktéž mohly mít vliv na produkci PHA, jelikož snížení aktivity glykolát oxidázy by způsobilo zvýšení toku Acetyl-CoA do metabolické dráhy syntézy PHA. Mutace #3 způsobující předčasné ukončení transkripce určitě má tento žádoucí efekt na aktivitu glykolát oxidázy. Efekt mutace #5 už není tak jasný, ale dá se předpokládat, že taktéž snižuje aktivitu glykolát oxidázy, jelikož zvýšení frekvence této mutace v populaci CN-Cu44 naznačuje její pozitivní selekci.

Mutace #11 by taktéž mohla ovlivnit produkci PHA, vzhledem však k její struktuře a okolnostem jejího stanovení je diskutabilní, zda jde o korektně stanovenou mutaci. Inhibice tohoto enzymu, či jeho úplné vyřazení, by snížením koncentrace Acetyl-CoA vedly k snížení rychlosti produkce PHA, a tedy k snížení odolnosti populace CN-Na41 v kontextu osmotického stresu. Pravděpodobnější vysvětlení tedy je chybné mapování u populací CN-Na41 a CN-wt.

5. Závěr

Cílem této práce bylo stanovení mutačních změn v populacích CN H16 adaptovaných na stresové podmínky v rámci evolučního experimentu Ing. Nováčkové, jejich následná analýza a stanovení případných souvislostí těchto mutačních změn s produkcí PHA.

Mutační změny byly stanoveny bioinformatickou analýzou dat z NGS CN H16 populací z experimentu – původní populace, populace adaptované na stres vyvolaný měďnatými ionty a populace adaptované na osmotický stres. Po ověření obecné kvality dat byly stanoveny mutační změny v adaptovaných populacích porovnáním readů ze všech tří populací vůči genomové sekvenci původního genomu se zohledněním mutací již přítomných na začátku experimentu. Následně byly dle charakteru stanovených mutací a na základě dostupných informací o relevantních úsecích genomu CN H16 vyhodnoceny souvislosti mezi mutačními změnami v populacích CN H16 a konkrétními stresovými podmínkami a produkcí PHA.

Kvůli omezeným informacím o úsecích genomu CN H16, v jakých byly mutace stanoveny, nebyly vyvozeny jednoznačné závěry o efektu těchto mutací na fenotyp CN H16. Dle teoretických předpokladů byl však na základě charakteru mutací, jejich frekvence v populacích a předpokládané funkci ovlivněných úseků genomu zaznamenán trend určité souvislosti stanovených mutací s adaptací na stresové podmínky experimentu. Jelikož CN H16 se dle teoretických předpokladů na podmínky vybrané při experimentu adaptuje mimo jiné produkcí PHA, byly očekávány mutační změny ovlivňující akumulaci PHA. Překvapivě jsme neobjevili žádné průkazné mutace právě v genech phaR, phaZ, phaP či v genech operonu phaCAB. Objevili jsme však mutace v genech, které syntézu PHA mohou ovlivňovat. Příkladem jsou tři mutace (#2, #7 a #8) u populace CN-Na41 (adaptované na osmotický stres) v genech souvisejících s regulací transkripce, přičemž citrátový cyklus je za osmotického stresu regulován především transkripcí. Dalšími příklady jsou mutace #3 u populace CN-Na41 a mutace #5 u CN-Cu44 v genech kódujících stejnou podjednotku glykolát oxidázy, jejíž aktivita může mít vliv na koncentraci Acetyl-CoA, která s rychlostí produkce PHA úzce souvisí.

6. Seznam použitých zdrojů

1. SPIEKERMANN, Patricia, Bernd H. A. REHM, Rainer KALSCHEUER, Dirk BAUMEISTER a A. STEINBÜCHEL. A sensitive, viable-colony staining method using Nile red for direct screening of bacteria that accumulate polyhydroxyalkanoic acids and other lipid storage compounds. *Archives of Microbiology* [online]. 1999, **171**(2), 73-80 [cit. 2020-03-02]. DOI: 10.1007/s002030050681. ISSN 0302-8933. Dostupné z: http://link.springer.com/10.1007/s002030050681
2. POHLMANN, Anne, Wolfgang Florian FRICKE, Frank REINECKE, et al. Genome sequence of the bioplastic-producing "Knallgas" bacterium Ralstonia eutropha H16. *Nature Biotechnology* [online]. 2006, **24**(10), 1257-1262 [cit. 2020-03-02]. DOI: 10.1038/nbt1244. ISSN 1087-0156. Dostupné z: http://www.nature.com/articles/nbt1244
3. LITTLE, Gareth T., Muhammad EHSAAN, Christian ARENAS-LÓPEZ, Kamran JAWED, Klaus WINZER, Katalin KOVACS, Nigel P. MINTON a David RASKO. Complete Genome Sequence of Cupriavidus necator H16 (DSM 428). *Microbiology Resource Announcements* [online]. 2019, **8**(37), e00814-19 [cit. 2020-03-16]. DOI: 10.1128/MRA.00814-19. ISSN 2576-098X. Dostupné z: http://genomea.asm.org/lookup/doi/10.1128/MRA.00814-19
4. MOTRO, Yair a Jacob MORAN-GILAD. Next-generation sequencing applications in clinical bacteriology. *Biomolecular Detection and Quantification* [online]. 2017, **14**, 1-6 [cit. 2020-03-02]. DOI: 10.1016/j.bdq.2017.10.002. ISSN 22147535. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S2214753517300050
5. NOVACKOVA, Ivana, Dan KUCERA, Jaromir PORIZKA, Iva PERNICOVA, Petr SEDLACEK, Martin KOLLER, Adriana KOVALCIK a Stanislav OBRUCA. Adaptation of Cupriavidus necator to levulinic acid for enhanced production of P(3HB-co-3HV) copolyesters. *Biochemical Engineering Journal* [online]. 2019, **151** [cit. 2020-03-02]. DOI: 10.1016/j.bej.2019.107350. ISSN 1369703X. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S1369703X19302876
6. Bacterial Genomes, In: *College of Agriculture and Life sciences* [online]. Ithaca, New York, US [cit. 2020-03-02]. Dostupné z: https://micro.cornell.edu/research/epulopiscium/bacterial-genomes/
7. BRISSON, Dustin, Dan DRECKTRAH, Christian H. EGGERS a D. Scott SAMUELS. Genetics of Borrelia burgdorferi. *Annual Review of Genetics* [online]. 2012, **46**(1), 515-536 [cit. 2020-03-02]. DOI: 10.1146/annurev-genet-011112-112140. ISSN 0066-4197. Dostupné z: http://www.annualreviews.org/doi/10.1146/annurev-genet-011112-112140
8. SHINTANI, Masaki, Zoe K. SANCHEZ a Kazuhide KIMBARA. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in Microbiology* [online]. 2015, **6** [cit. 2020-03-02]. DOI: 10.3389/fmicb.2015.00242. ISSN 1664-302X. Dostupné z: http://www.frontiersin.org/Evolutionary_and_Genomic_Microbiology/10.3389/fmicb.2015.00242/abstract
9. KOONIN, Eugene V. a Artem S. NOVOZHILOV. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* [online]. 2009, **61**(2), 99-111 [cit. 2020-03-03]. DOI: 10.1002/iub.146. ISSN 15216543. Dostupné z: http://doi.wiley.com/10.1002/iub.146
10. KOONIN, Eugene V. Evolution of genome architecture. *The International Journal of Biochemistry & Cell Biology* [online]. 2009, **41**(2), 298-306 [cit. 2020-03-03]. DOI: 10.1016/j.biocel.2008.09.015. ISSN 13572725. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S1357272508003907

11. ORSINI, Massimiliano, Gianmauro CUCCURU, Paolo UVA a Giorgio FOTIA. Bacterial Genomic Data Analysis in the Next-Generation Sequencing Era. CARUGO, Oliviero a Frank EISENHABER, ed. *Data Mining Techniques for the Life Sciences* [online]. New York, NY: Springer New York, 2016, 2016-04-27, s. 407-422 [cit. 2020-03-03]. Methods in Molecular Biology. DOI: 10.1007/978-1-4939-3572-7_21. ISBN 978-1-4939-3570-3. Dostupné z: http://link.springer.com/10.1007/978-1-4939-3572-7_21
12. Watford S, Warrington SJ. Bacterial DNA Mutations. [Updated 2019 Apr 27]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 03.03.2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459274/
13. O'DONNELL, M., L. LANGSTON a B. STILLMAN. Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya. *Cold Spring Harbor Perspectives in Biology* [online]. 2013, **5**(7), a010108-a010108 [cit. 2020-03-03]. DOI: 10.1101/cshperspect.a010108. ISSN 1943-0264. Dostupné z: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a010108
14. DEMAIN, Arnold L. a Jose L. ADRIO. Strain improvement for production of pharmaceuticals and other microbial metabolites by fermentation. PETERSEN, Frank a René AMSTUTZ, ed. *Natural Compounds as Drugs Volume I* [online]. Basel: Birkhäuser Basel, 2008, s. 251-289 [cit. 2020-03-03]. Progress in Drug Research. DOI: 10.1007/978-3-7643-8117-2_7. ISBN 978-3-7643-8098-4. Dostupné z: http://link.springer.com/10.1007/978-3-7643-8117-2_7
15. CALVIN, N M a P C HANAWALT. High-efficiency transformation of bacterial cells by electroporation. *Journal of Bacteriology* [online]. 1988, **170**(6), 2796-2801 [cit. 2020-03-03]. DOI: 10.1128/JB.170.6.2796-2801.1988. ISSN 0021-9193. Dostupné z: https://JB.asm.org/content/170/6/2796
16. Neumann E, Schaefer-Ridder M, Wang Y, Hofschneider PH. Gene transfer into mouse lyoma cells by electroporation in high electric fields. *EMBO J*. 1982;1(7):841–845. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC553119/
17. WALSH, Gary. Therapeutic insulins and their large-scale manufacture. *Applied Microbiology and Biotechnology* [online]. 2005, **67**(2), 151-159 [cit. 2020-03-03]. DOI: 10.1007/s00253-004-1809-x. ISSN 0175-7598. Dostupné z: http://link.springer.com/10.1007/s00253-004-1809-x
18. BAESHEN, Nabih A, Mohammed N BAESHEN, Abdullah SHEIKH, Roop S BORA, Mohamed Morsi M AHMED, Hassan A I RAMADAN, Kulvinder Singh SAINI a Elrashdy M REDWAN. Cell factories for insulin production. *Microbial Cell Factories* [online]. 2014, **13**(1) [cit. 2020-03-03]. DOI: 10.1186/s12934-014-0141-0. ISSN 1475-2859. Dostupné z: http://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-014-0141-0
19. Restrictions on Genetically Modified Organisms, *The Library of Congress* [online]. Březen 2014 [cit. 2020-03-03]. Dostupné z: https://www.loc.gov/law/help/restrictions-on-gmos/
20. HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. 2016, **107**(1), 1-8 [cit. 2020-03-03]. DOI: 10.1016/j.ygeno.2015.11.003. ISSN 08887543. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0888754315300410
21. SANGER, F., G. M. AIR, B. G. BARRELL, et al. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* [online]. 1977, **265**(5596), 687-695 [cit. 2020-03-03]. DOI: 10.1038/265687a0. ISSN 0028-0836. Dostupné z: http://www.nature.com/articles/265687a0
22. HOOD, Leroy a Lee ROWEN. The human genome project: big science transforms biology and medicine. *Genome Medicine* [online]. 2013, **5**(9) [cit. 2020-03-03]. DOI:

10.1186/gm483. ISSN 1756-994X. Dostupné z: http://genomemedicine.biomedcentral.com/articles/10.1186/gm483
23. Six Years After Acquisition, Roche Quietly Shutters 454, *Bio IT World* [online]. 2013 [cit. 2020-03-03]. Dostupné z: http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shutters-454.html
24. WETTERSTRAND, M.S., Kris A., The Cost of Sequencing a Human Genome. *National Human Genome Research Institute* [online]. Bethesda, Maryland, October 30, 2019 [cit. 2020-03-03]. Dostupné z: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost
25. MOROZOVA, Olena a Marco A. MARRA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* [online]. 2008, **92**(5), 255-264 [cit. 2020-03-08]. DOI: 10.1016/j.ygeno.2008.07.001. ISSN 08887543. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0888754308001651
26. MILLER, Jason R., Sergey KOREN a Granger SUTTON. Assembly algorithms for next-generation sequencing data. *Genomics* [online]. 2010, **95**(6), 315-327 [cit. 2020-03-08]. DOI: 10.1016/j.ygeno.2010.03.001. ISSN 08887543. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0888754310000492
27. PFEIFER, S P. From next-generation resequencing reads to a high-quality variant data set. *Heredity* [online]. 2017, **118**(2), 111-124 [cit. 2020-03-08]. DOI: 10.1038/hdy.2016.102. ISSN 0018-067X. Dostupné z: http://www.nature.com/articles/hdy2016102
28. AMBARDAR, Sheetal, Rikita GUPTA, Deepika TRAKROO, Rup LAL a Jyoti VAKHLU. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology* [online]. 2016, **56**(4), 394-404 [cit. 2020-03-08]. DOI: 10.1007/s12088-016-0606-4. ISSN 0046-8991. Dostupné z: http://link.springer.com/10.1007/s12088-016-0606-4
29. File Format Guide, 1988. *NCBI: National Center for Biology Information* [online]. Bethesda, Maryland U.S. [cit. 2020-03-13]. Dostupné z: https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/#introduction
30. DJAKOW, Jana, Lenka KRAMNÁ, Lenka DUŠÁTKOVÁ, Jiří UHLÍK, Juha-Pekka PURSIHEIMO, Tamara SVOBODOVÁ, Petr POHUNEK a Ondřej CINEK. An effective combination of sanger and next generation sequencing in diagnostics of primary ciliary dyskinesia. *Pediatric Pulmonology* [online]. 2016, **51**(5), 498-509 [cit. 2020-03-13]. DOI: 10.1002/ppul.23261. ISSN 87556863. Dostupné z: http://doi.wiley.com/10.1002/ppul.23261
31. PRATAS, Diogo, Morteza HOSSEINI a Armando J. PINHO. Cryfa: A Tool to Compact and Encrypt FASTA Files. FDEZ-RIVEROLA, Florentino, Mohd Saberi MOHAMAD, Miguel ROCHA, Juan F. DE PAZ a Tiago PINTO, ed. *11th International Conference on Practical Applications of Computational Biology & Bioinformatics* [online]. Cham: Springer International Publishing, 2017, 2017-06-21, s. 305-312 [cit. 2020-03-13]. Advances in Intelligent Systems and Computing. DOI: 10.1007/978-3-319-60816-7_37. ISBN 978-3-319-60815-0. Dostupné z: http://link.springer.com/10.1007/978-3-319-60816-7_37
32. The Variant Call Format (VCF) Version 4.1 Specification, 2008. *GitHub* [online]. San Francisco, California U.S., 8.7.2019 [cit. 2020-03-13]. Dostupné z: https://samtools.github.io/hts-specs/VCFv4.1.pdf
33. EDWARDS, David J a Kathryn E HOLT. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation* [online]. 2013, **3**(1) [cit. 2020-03-14]. DOI: 10.1186/2042-5783-3-2. ISSN

2042-5783. Dostupné z: https://microbialinformaticsj.biomedcentral.com/articles/10.1186/2042-5783-3-2
34. Definition of compiler, 1981. *PC Magazine: PCMag* [online]. New York, U.S.: ZIFF DAVIS, LLC. PCMAG DIGITAL GROUP [cit. 2020-03-14]. Dostupné z: https://www.pcmag.com/
35. AFGAN, Enis, Dannon BAKER, Bérénice BATUT, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research* [online]. 2018, **46**(W1), W537-W544 [cit. 2020-03-14]. DOI: 10.1093/nar/gky379. ISSN 0305-1048. Dostupné z: https://academic.oup.com/nar/article/46/W1/W537/5001157
36. BLANKENBERG, Daniel, Gregory Von KUSTER, Nathaniel CORAOR, Guruprasad ANANDA, Ross LAZARUS, Mary MANGAN, Anton NEKRUTENKO a James TAYLOR. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology* [online]. 2010, **89**(1) [cit. 2020-03-14]. DOI: 10.1002/0471142727.mb1910s89. ISSN 1934-3639. Dostupné z: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb1910s89
37. FASTQC: A quality control tool for high throughput sequence data., *Babraham Bioinformatics: Babraham Institute* [online]. Babraham Research Campus, Babraham, Cambridgeshire, UK: University of Cambridge, 8.1.2019 [cit. 2020-03-14]. Dostupné z: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
38. CHEN, Shifu, Yanqing ZHOU, Yaru CHEN a Jia GU. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* [online]. 2018, **34**(17), i884-i890 [cit. 2020-03-14]. DOI: 10.1093/bioinformatics/bty560. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article/34/17/i884/5093234
39. LI, H., B. HANDSAKER, A. WYSOKER, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [online]. 2009, **25**(16), 2078-2079 [cit. 2020-03-14]. DOI: 10.1093/bioinformatics/btp352. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp352
40. KOBOLDT, D. C., Q. ZHANG, D. E. LARSON, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* [online]. 2012, **22**(3), 568-576 [cit. 2020-03-30]. DOI: 10.1101/gr.129684.111. ISSN 1088-9051. Dostupné z: http://genome.cshlp.org/cgi/doi/10.1101/gr.129684.111
41. ALHAMDOOSH, Monther, Milica NG, Nicholas J. WILSON, Julie M. SHERIDAN, Huy HUYNH, Michael J. WILSON a Matthew E. RITCHIE. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics* [online]. [cit. 2020-03-14]. DOI: 10.1093/bioinformatics/btw623. ISSN 1367-4803. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw623
42. KUTRALAM-MUNIASAMY, Gurusamy a Fermín PERÉZ-GUEVARA. Genome characteristics dictate poly-R-(3)-hydroxyalkanoate production in Cupriavidus necator H16. *World Journal of Microbiology and Biotechnology* [online]. 2018, **34**(6) [cit. 2020-03-16]. DOI: 10.1007/s11274-018-2460-5. ISSN 0959-3993. Dostupné z: http://link.springer.com/10.1007/s11274-018-2460-5
43. MADISON, L. L., a G. W. HUISMAN. Metabolic engineering of poly(3-hydroxyalkanoates): from DNA to plastic. *Microbiology and molecular biology reviews : MMBR* vol. 63,1 (1999): 21-53. Dostupné z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC98956/

44. SEDLACEK, Petr, Eva SLANINOVA, Martin KOLLER, Jana NEBESAROVA, Ivana MAROVA, Vladislav KRZYZANEK a Stanislav OBRUCA. PHA granules help bacterial cells to preserve cell integrity when exposed to sudden osmotic imbalances. *New Biotechnology* [online]. 2019, **49**, 129-136 [cit. 2020-03-16]. DOI: 10.1016/j.nbt.2018.10.005. ISSN 18716784. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S1871678418305119
45. *Songklanakarin J. Sci. Technol.: Effect of co-substrate on production of poly-β-hydroxybutyrate (PHB) and copolymer PHBV from newly identified mutant Rhodobacter sphaeroides U7 cultivated under aerobic-dark condition* [online], 2007,29(4). [cit. 2020-03-19]. Dostupné z: https://www.researchgate.net/publication/26484981_Effect_of_co-substrate_on_production_of_poly-b-_hydroxybutyrate_PHB_and_copolymer_PHBV_from_newly_identified_mutant_Rhodobacter_sphaeroides_U7_cultivated_under_aerobic-dark_condition
46. OBRUCA, Stanislav, Pavla BENESOVA, Jana OBORNA a Ivana MAROVA. Application of protease-hydrolyzed whey as a complex nitrogen source to increase poly(3-hydroxybutyrate) production from oils by Cupriavidus necator. *Biotechnology Letters* [online]. 2014, **36**(4), 775-781 [cit. 2020-03-16]. DOI: 10.1007/s10529-013-1407-z. ISSN 0141-5492. Dostupné z: http://link.springer.com/10.1007/s10529-013-1407-z
47. CAI, Shuangfeng, Lei CAI, Dahe ZHAO, Guiming LIU, Jing HAN, Jian ZHOU, Hua XIANG a M. KIVISAAR. A Novel DNA-Binding Protein, PhaR, Plays a Central Role in the Regulation of Polyhydroxyalkanoate Accumulation and Granule Formation in the Haloarchaeon Haloferax mediterranei. *Applied and Environmental Microbiology* [online]. 2014, **81**(1), 373-385 [cit. 2020-03-17]. DOI: 10.1128/AEM.02878-14. ISSN 0099-2240. Dostupné z: http://aem.asm.org/lookup/doi/10.1128/AEM.02878-14
48. YORK, G. M., B. H. JUNKER, J. STUBBE a A. J. SINSKEY. Accumulation of the PhaP Phasin of Ralstonia eutropha Is Dependent on Production of Polyhydroxybutyrate in Cells. *Journal of Bacteriology* [online]. 2001, **183**(14), 4217-4226 [cit. 2020-03-17]. DOI: 10.1128/JB.183.14.4217-4226.2001. ISSN 0021-9193. Dostupné z: http://jb.asm.org/cgi/doi/10.1128/JB.183.14.4217-4226.2001
49. KUCHTA, Kenny, Lifeng CHI, Harald FUCHS, Markus PÖTTER a Alexander STEINBÜCHEL. Studies on the Influence of Phasins on Accumulation and Degradation of PHB and Nanostructure of PHB Granules in Ralstonia eutropha H16. *Biomacromolecules* [online]. 2007, **8**(2), 657-662 [cit. 2020-03-17]. DOI: 10.1021/bm060912e. ISSN 1525-7797. Dostupné z: https://pubs.acs.org/doi/10.1021/bm060912e
50. JENDROSSEK, D., A. SCHIRMER a H. G. SCHLEGEL. Biodegradation of polyhydroxyalkanoic acids. *Applied Microbiology and Biotechnology* [online]. 1996, **46**(5-6), 451-463 [cit. 2020-03-17]. DOI: 10.1007/s002530050844. ISSN 0175-7598. Dostupné z: http://link.springer.com/10.1007/s002530050844
51. KHARE, Ekta, Jyotsana CHOPRA a Naveen Kumar ARORA. Screening for MCL-PHA-Producing Fluorescent Pseudomonads and Comparison of MCL-PHA Production Under Iso-osmotic Conditions Induced by PEG and NaCl. *Current Microbiology* [online]. 2014, **68**(4), 457-462 [cit. 2020-03-19]. DOI: 10.1007/s00284-013-0497-0. ISSN 0343-8651. Dostupné z: http://link.springer.com/10.1007/s00284-013-0497-0
52. WANG, Yayi, Zhongjia REN, Fan JIANG, Junjun GENG, Weitao HE a Jian YANG. Effect of copper ion on the anaerobic and aerobic metabolism of phosphorus-accumulating organisms linked to intracellular storage compounds. *Journal of Hazardous Materials* [online]. 2011, **186**(1), 313-319 [cit. 2020-03-31]. DOI:

10.1016/j.jhazmat.2010.11.007. ISSN 03043894. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0304389410014135
53. *DNALink: Your link to precision genomics* [online], 2000. Seoul, Republic of South Korea [cit. 2020-03-20]. Dostupné z: http://www.dnalink.com/english/
54. TRIVEDI, Urmi H., TimothÃ©e CÃ©ZARD, Stephen BRIDGETT, Anna MONTAZAM, Jenna NICHOLS, Mark BLAXTER a Karim GHARBI. Quality control of next-generation sequencing data without a reference. *Frontiers in Genetics* [online]. 2014, **5** [cit. 2020-03-21]. DOI: 10.3389/fgene.2014.00111. ISSN 1664-8021. Dostupné z: http://journal.frontiersin.org/article/10.3389/fgene.2014.00111/abstract
55. SEEMANN, Torsten a Simon GLADMAN, 2012. *Display summary statistics for a fasta file*. Victorian Bioinformatics Consortium.
56. DARLING, A. C.E. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* [online]. 2004, **14**(7), 1394-1403 [cit. 2020-03-24]. DOI: 10.1101/gr.2289704. ISSN 1088-9051. Dostupné z: http://www.genome.org/cgi/doi/10.1101/gr.2289704
57. Mauve: Multiple Genome Alignment, *The Darling lab: computational (meta)genomics* [online]. University of Technology Sydney [cit. 2020-04-17]. Dostupné z: http://darlinglab.org/mauve/mauve.html
58. CABANETTES, Floréal a Christophe KLOPP. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* [online]. 2018, **6** [cit. 2020-03-24]. DOI: 10.7717/peerj.4958. ISSN 2167-8359. Dostupné z: https://peerj.com/articles/4958
59. KEEL, Brittney N. a Warren M. SNELLING. Comparison of Burrows-Wheeler Transform-Based Mapping Algorithms Used in High-Throughput Whole-Genome Sequencing: Application to Illumina Data for Livestock Genomes1. *Frontiers in Genetics* [online]. 2018, **9** [cit. 2020-03-27]. DOI: 10.3389/fgene.2018.00035. ISSN 1664-8021. Dostupné z: http://journal.frontiersin.org/article/10.3389/fgene.2018.00035/full
60. KISHIKAWA, Toshihiro, Yukihide MOMOZAWA, Takeshi OZEKI, Taisei MUSHIRODA, Hidenori INOHARA, Yoichiro KAMATANI, Michiaki KUBO a Yukinori OKADA. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Scientific Reports* [online]. 2019, **9**(1) [cit. 2020-03-22]. DOI: 10.1038/s41598-018-38346-0. ISSN 2045-2322. Dostupné z: http://www.nature.com/articles/s41598-018-38346-0
61. NAKAMURA, Kensuke, Taku OSHIMA, Takuya MORIMOTO, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* [online]. 2011, **39**(13), e90-e90 [cit. 2020-03-22]. DOI: 10.1093/nar/gkr344. ISSN 1362-4962. Dostupné z: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr344
62. WICK, Ryan R., Louise M. JUDD, Claire L. GORRIE, Kathryn E. HOLT a Adam M. PHILLIPPY. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* [online]. 2017, **13**(6) [cit. 2020-04-18]. DOI: 10.1371/journal.pcbi.1005595. ISSN 1553-7358. Dostupné z: https://dx.plos.org/10.1371/journal.pcbi.1005595
63. ALTSCHUL, Stephen F., Warren GISH, Webb MILLER, Eugene W. MYERS a David J. LIPMAN. Basic local alignment search tool. *Journal of Molecular Biology* [online]. 1990, **215**(3), 403-410 [cit. 2020-04-19]. DOI: 10.1016/S0022-2836(05)80360-2. ISSN 00222836. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602

64. *NCBI: National Center for Biotechnology Information* [online], 8600 Rockville Pike, Bethesda MD, 20894 USA: U.S. National Library of Medicine [cit. 2020-04-19]. Dostupné z: https://www.ncbi.nlm.nih.gov/
65. LI, Heng. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* [online]. 2014, **30**(20), 2843-2851 [cit. 2020-04-01]. DOI: 10.1093/bioinformatics/btu356. ISSN 1460-2059. Dostupné z: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu356
66. DE VEGA, Jose, 2017. *SNP Calling vs. sequencing coverage: Cost-effective approaches for variant calling and analysis in complex plants*. Earlham Institute, United Kingdom. Dostupné také z: https://www.fgua.es/wp-content/uploads/2017/11/dVegaS3.pdf
67. DOUGLASS, Alexander P., Caoimhe E. O'BRIEN, Benjamin OFFEI, Aisling Y. COUGHLAN, Raúl A. ORTIZ-MERINO, Geraldine BUTLER, Kevin P. BYRNE a Kenneth H. WOLFE. Coverage-Versus-Length Plots, a Simple Quality Control Step for de Novo Yeast Genome Sequence Assemblies. *G3: Genes|Genomes|Genetics* [online]. [cit. 2020-04-02]. DOI: 10.1534/g3.118.200745. ISSN 2160-1836. Dostupné z: http://g3journal.org/lookup/doi/10.1534/g3.118.200745
68. YOUNG, GregoryB., DonaldL. JACK, DouglasW. SMITH a MiltonH. SAIER. The amino acid/auxin: proton symport permease family1The accompanying review paper 'Phylogenetic characterization of novel transport protein families revealed by genome analysis' by M.H. Saier Jr. et al. will be published in Biochim. Biophys. Acta, Vol. 1422/1, February 1999 issue.1. *Biochimica et Biophysica Acta (BBA) - Biomembranes* [online]. 1999, **1415**(2), 306-322 [cit. 2020-04-05]. DOI: 10.1016/S0005-2736(98)00196-5. ISSN 00052736. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0005273698001965
69. CSONKA, L S, 1989. Physiological and genetic responses of bacteria to osmotic stress. *Microbiology Reviews: Now published as Microbiology and Molecular Biology Reviews*. Washington, D.C., U.S.: American society for microbiology, **53**(1), 121-147. PMID: 2651863. Dostupné také z: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC372720/
70. DAS, Debanu, Nick V. GRISHIN, Abhinav KUMAR, et al. The structure of the first representative of Pfam family PF09836 reveals a two-domain organization and suggests involvement in transcriptional regulation. *Acta Crystallographica Section F Structural Biology and Crystallization Communications* [online]. 2010, **66**(10), 1174-1181 [cit. 2020-04-05]. DOI: 10.1107/S1744309109022672. ISSN 1744-3091. Dostupné z: http://scripts.iucr.org/cgi-bin/paper?S1744309109022672
71. BARTHOLOMÄUS, Alexander, Ivan FEDYUNIN, Peter FEIST, Celine SIN, Gong ZHANG, Angelo VALLERIANI a Zoya IGNATOVA. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* [online]. 2016, **374**(2063) [cit. 2020-04-05]. DOI: 10.1098/rsta.2015.0069. ISSN 1364-503X. Dostupné z: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0069
72. MADDOCKS, Sarah E. a Petra C. F. OYSTON. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* [online]. 2008, **154**(12), 3609-3623 [cit. 2020-04-05]. DOI: 10.1099/mic.0.2008/022772-0.

ISSN 1350-0872. Dostupné z: https://www.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.2008/022772-0
73. OSORIO, CR, MD COLLINS, JL ROMALDE a AE TORANZO. Characterization of the 23S and 5S rRNA genes and 23S-5S intergenic spacer region (ITS-2) of Photobacterium damselae. *Diseases of Aquatic Organisms* [online]. 2004, **61**, 33-39 [cit. 2020-04-06]. DOI: 10.3354/dao061033. ISSN 0177-5103. Dostupné z: http://www.int-res.com/abstracts/dao/v61/n1-2/p33-39/
74. TILSALA-TIMISJÄRVI, Anu a Tapani ALATOSSAVA. Characterization of the 16S–23S and 23S–5S rRNA intergenic spacer regions of dairy propionibacteria and their identification with species-specific primers by PCR. *International Journal of Food Microbiology* [online]. 2001, **68**(1-2), 45-52 [cit. 2020-04-06]. DOI: 10.1016/S0168-1605(01)00462-7. ISSN 01681605. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0168160501004627
75. KLITGAARD, Rasmus N., Eleni NTOKOU, Katrine NØRGAARD, Daniel BILTOFT, Lykke H. HANSEN, Nicolai M. TRÆDHOLM, Jacob KONGSTED a Birte VESTER. Mutations in the Bacterial Ribosomal Protein L3 and Their Association with Antibiotic Resistance. *Antimicrobial Agents and Chemotherapy* [online]. 2015, **59**(6), 3518-3528 [cit. 2020-04-06]. DOI: 10.1128/AAC.00179-15. ISSN 0066-4804. Dostupné z: http://aac.asm.org/lookup/doi/10.1128/AAC.00179-15
76. LONG, K. S., C. MUNCK, T. M. B. ANDERSEN, M. A. SCHAUB, S. N. HOBBIE, E. C. BOTTGER a B. VESTER. Mutations in 23S rRNA at the Peptidyl Transferase Center and Their Relationship to Linezolid Binding and Cross-Resistance. *Antimicrobial Agents and Chemotherapy* [online]. 2010, **54**(11), 4705-4713 [cit. 2020-04-06]. DOI: 10.1128/AAC.00644-10. ISSN 0066-4804. Dostupné z: http://aac.asm.org/cgi/doi/10.1128/AAC.00644-10
77. TAMÁS, Markus, Sandeep SHARMA, Sebastian IBSTEDT, Therese JACOBSON a Philipp CHRISTEN. Heavy Metals and Metalloids As a Cause for Protein Misfolding and Aggregation. *Biomolecules* [online]. 2014, **4**(1), 252-267 [cit. 2020-04-06]. DOI: 10.3390/biom4010252. ISSN 2218-273X. Dostupné z: http://www.mdpi.com/2218-273X/4/1/252
78. SAMANTA, D., D. MUKHOPADHYAY, S. CHOWDHURY, et al. Protein Folding by Domain V of Escherichia coli 23S rRNA: Specificity of RNA-Protein Interactions. *Journal of Bacteriology* [online]. 2008, **190**(9), 3344-3352 [cit. 2020-04-06]. DOI: 10.1128/JB.01800-07. ISSN 0021-9193. Dostupné z: http://jb.asm.org/cgi/doi/10.1128/JB.01800-07
79. LEE, J.-N., H.-D. SHIN a Y.-H. LEE. Metabolic Engineering of Pentose Phosphate Pathway in Ralstonia eutropha for Enhanced Biosynthesis of Poly-β-hydroxybutyrate. *Biotechnology Progress* [online]. 2003, **19**(5), 1444-1449 [cit. 2020-04-06]. DOI: 10.1021/bp034060v. ISSN 8756-7938. Dostupné z: http://doi.wiley.com/10.1021/bp034060v

7. Seznam příloh

1. DNALink: CN-wt Analysis_Report.pdf
2. DNALink: CN-Cu44 Analysis_Report.pdf
3. DNALink: CN-Na41 Analysis_Report.pdf
4. FASTQC: CN-wt R1 Report.html
5. FASTQC: CN-wt R2 Report.html
6. FASTQC: CN-Cu44 R1 Report.html
7. FASTQC: CN-Cu44 R2 Report.html
8. FASTQC: CN-Na41 R1 Report.html
9. FASTQC: CN-Na41 R2 Report.html
10. VarScan 2 Report.vcf
11. Seznam 139 mutací.xlsx
12. NCBI-Blast report - phaCAB + phaR porovnáno s CN-wt.txt
13. NCBI-Blast report - phaP vs CN-wt.txt